VEGETATION CLASSIFICATION AND SURVEY

International Association for Vegetation Science (IAVS)

Vegetation Classification and Survey 6: 99–127
doi: 10.3897/VCS.140466

VCS METHODS ECOINFORMATICS

TDV-optimization: A novel numerical method for phytosociological tabulation

Tiago Monteiro-Henriques[1]

1 *Global Change and Conservation Lab (GCC), Faculty of Biological and Environmental Sciences, University of Helsinki, Helsinki, Finland*

Corresponding author: Tiago Monteiro-Henriques (tmh@isa.ulisboa.pt)

Academic editor: David W. Roberts ♦ **Linguistic editor:** Lynda Weeks

Received 29 October 2024 ♦ **Accepted** 1 March 2025 ♦ **Published** 17 April 2025

Abstract

I present the Total Differential Value (TDV), an index designed for vegetation analysis based on the operational concept of differential species, as classically illustrated by Heinz Ellenberg and Dieter Mueller-Dombois. Given a phytosociological table and a grouping of its relevés, TDV is obtained by averaging the Differential Value (DiffVal) for each species in the table. DiffVal, grounded in combinatorial-discrete mathematics, quantifies the differential power of a species. The novelty of this approach lies in its distinction between two types of species absences: (i) absences from some relevés within a group and (ii) absences from all relevés representing a group. By leveraging the distribution of species absences among groups, this method effectively quantifies the differential power and distinguishes differential from non-differential species. I illustrate the computation of DiffVal and TDV and show that, because only differential species contribute to TDV, it reflects the strength of the differential species patterns in a classified table. TDV can be optimized (TDV-optimization), providing partitions of relevés. I demonstrate TDV-optimization using both an artificial and a well-known real-world data set. Key features of this method include its ability to identify patterns very closely resembling manual phytosociological tabulation and to detect reticulate patterns. TDV-optimization may lead to partitions where outlier or extreme relevés are isolated in groups; however, enforcing a minimum group size can highlight partitions with more balanced group sizes. An R package is now available, implementing DiffVal and TDV calculation as well as TDV-optimization.

Abbreviations: DiffVal = Differential Value; EllPar = Ellenberg's partition into three groups; NDR = number of discrepant relevés; NSS = number of significant species; TDV = Total Differential Value.

Keywords

biclustering, block structure, differential species, optimization, patterns, phytosociological tabulation, tabular classification, TDV-optimization, vegetation

Introduction

Vegetation is a complex phenomenon driven by many deeply interacting factors. As bottom-up mechanistic approaches to vegetation are often hindered by this complexity, physiognomic-floristic patterns are sought as a means to simplify and handle this phenomenon (Armbruster 1995; Austin 2013).

Vegetation classification is the process of identifying, describing and interrelating vegetation units using vegetation relevés (see, e.g., Peet and Roberts 2013, p. 29). In specialized fields such as data science, machine learning, or pattern recognition, the term *classification* typically refers to the assignment of new individuals to existing classes, while *clustering* or *partitioning* refers to the process of identifying those classes (Bishop 2006; Witten 2009;

PENSOFT

Copyright *Tiago Monteiro-Henriques*. This is an open access article distributed under the terms of the Creative Commons Attribution License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Drummond 2011; Han et al. 2012; Bouveyron et al. 2019). In vegetation science, however, 'classification' is used more broadly to encompass both the establishment of classes and the assignment of objects to those classes.

Early in its history, vegetation classification was conducted solely by manipulating matrices of vegetation samples (relevés), as later described in detail by Ellenberg (1956) and translated to English by Mueller-Dombois and Ellenberg (1974). This manipulation forms the basis of Braun-Blanquet's approach to vegetation, with the main purpose being the partitioning of relevés. In this process, the researcher seeks to divide the relevé data set into groups (subsets of relevés) that can be distinguished floristically (i.e., by subsets of the occurring species). The process is potentially highly subjective and particularly complex, even for moderately sized tables. The number of groups is expected to emerge naturally during the manipulation, adding further complexity. This technique is called *tabulation* (Poore 1955), *tablework* or *table sorting* (Češka and Roemer 1971), *tabular sorting* (Dale and Quadraccia 1973; Dale and Webb 1975), or *tabular classification* (Wildi 1989). The output of the technique is referred to as a *differentiated table* (Ellenberg 1956; Spatz 1972; Mueller-Dombois and Ellenberg 1974), or colloquially as a *classified*, *ordered*, *rearranged* or *sorted table* – simply put, a *tabulation*. After obtaining the groups, it is common to make fine adjustments to the species sorting, such as highlighting species considered characteristic of already known higher-ranked syntaxa, primarily to improve the final display of the table. Thus, part of this fine-tuning relies on information not contained in the relevé matrix (e.g., the current syntaxonomic system) and should be treated separately from clustering/partitioning, to which it is irrelevant.

As described in Ellenberg (1956), Whittaker (1962), and Mueller-Dombois and Ellenberg (1974), the first step of the tabulation is the recognition of species groups, known as *differential species*. These groups are identified based on their co-occurrence in a certain number of relevés, while being absent from others. Henceforth, I refer to these as *patterns of differential species*.

The concept of differential species dates back over 100 years (Braun-Blanquet 1925) and forms the foundation of the tabulation technique, as well as the delimitation of vegetation types (for a detailed analysis of the concept and its changes over the years, see Appendix 1). Mueller-Dombois and Ellenberg (1974, p. 178) describe "differential (or differentiating) species" as those found among the "species with restricted distributions or amplitudes in the relevés under comparison". These authors operationalized the concept in its broadest sense, identifying species that are present in one or more groups but absent from others (such species cannot occur in all the emerging groups; they must be absent from at least one group to be considered differential). In this work, I adopt this definition, further exploring the notion of *exclusiveness* as presented and illustrated by Mueller-Dombois and Ellenberg (1974). It is worth noting that their definition differs slightly from the one given by Braun-Blanquet (1925, p. 134; 1932, p. 59), a distinction thoroughly examined in Appendix 1.

The concept of differential species is closely linked to two other key concepts from the Braun-Blanquet school: (i) *characteristic species* and (ii) *fidelity*. These concepts have also endured changes over the years. For a detailed discussion, see Appendix 1.

During the first half of the 20th century, researchers discussed the feasibility of vegetation classification based on physiognomic-floristic patterns, emphasizing the need for objective sampling and classification methods (Goodall 1953; Whittaker 1962). A vast array of numerical methods for sampling vegetation and identifying patterns emerged in the latter half of the 20th century (e.g., Whittaker 1962; Mueller-Dombois and Ellenberg 1974; Mucina and van der Maarel 1989) and continues into the 21st century (e.g., Peet and Roberts 2013). Most methods aimed to replicate or approximate the tabulation technique, yet with diverse approaches and results. With such an abundance of methods, it seems unlikely that the current status quo can still benefit from further contributions.

The results of numerical methods are often unsatisfactory for vegetation scientists (Roberts 2015), as they fail to replicate traditional tabular outputs (Mucina 1997). This may stem from the inappropriateness of certain algorithms for vegetation classification (Schmidtlein et al. 2010). Table 1 synthesizes the most relevant clustering/partitioning approaches used in vegetation analysis (including the one presented in this work). The table highlights the grouping approach (or model) and the space in which the groups are sought. The inherent characteristics of the model and space can influence the method's suitability for the analysis objectives, as well as its consistency with theoretical models, expert knowledge or even ground truth. To address this, some key characteristics that can impact consistency are presented in Table 1 as questions. As the present work focuses on phytosociological tabulation as outlined in Mueller-Dombois and Ellenberg (1974), I highlight in green those answers that satisfy the tabulation mechanics and align with the concept of differential species, and I highlight in red those that do not. Four topics are addressed:

(1) Hierarchical structure

Algorithms that produce hierarchical structures are constrained by nested fusion/division (Peet and Roberts 2013), which can hinder optimality for certain clustering criteria. Additionally, *this structure is not part of the tabulation procedure*.

(2) Geometric search

Schmidtlein et al. (2010) noted that expert groupings may be irregular in attribute space, while most numerical methods assume spherical or regular geometries. Regardless of

Table 1. Clustering/partitioning approaches commonly used in vegetation analysis. Each approach is characterized by the models on which the grouping is based, the space in which the groups are sought, and other relevant characteristics that may hinder consistency with theoretical models, reality or expert knowledge.

Name	Reference	Model: Grouping is based on...							Space: Groups are searched in...						Relevant characteristics			
		...clusters centre (centroid, medoid, etc.)	...an agglomerative strategy	...a divisive strategy	...density	...biclusters (subsets)	...graph theory	...heuristics[1]	...metaheuristcs[1]	...the data space (as a Euclidean space)	...pairwise distance/dissimilarity space	...an ordination space, plane or axis	...the data space (the original table)	...a graph	...the solution/partition space	Returns a hierarchical structure?	Search is geometric?	Species are considered equally informative?	Ensures exclusiveness of informative species?
Agglomerative hierarchical clustering	Sokal and Sneath (1963); Ward (1963)		■								■					Y	Y	Y	---
k-means	MacQueen (1967)	■						■		■						N	Y	Y	---
---	Češka and Roemer (1971)[2]					■		■					■			N	N	N	O[4]
TABORD	van der Maarel et al. (1978)	■	■					■		□	□					N	Y	Y	---
TWINSPAN	Hill (1979)	■		■				■				■				Y	Y	P[5]	N
COMPCLUS	Gauch (1980)				■			■		■						N	N	Y	---
ALOC	Belbin (1987)	■	■					■		□	□					N	Y	Y	---
PAM	Kaufman and Rousseeuw (1990)	■						■		■						N	Y	Y	---
DIANA	Kaufman and Rousseeuw (1990)			■							■					Y	Y	Y	---
---	Podani and Feoli (1991)					■			■						■	N	N	Y	---
ESPRESSO	Bruelheide and Flintrop (1994)[2]					■		■					■			N	N	N	N
COCKTAIL	Bruelheide (2000) [2]					■		■					■			N	N	N	N
ISOPAM	Schmidtlein et al. (2010)	■		□				■			■[6]					O	P[7]	Y[8]	N
---	Thiergart et al. (2014)						■							■[9]		N	N	P[10]	N
---	Mikolajczak et al. (2015)[3]						■							■[11]		N	N	P[12]	N
OPTPART	Roberts (2015)								■						■	N	N	Y	---
OPTSIL	Roberts (2015)								■						■	N	N	Y	---
REMOS	Lengyel et al. (2021)							■							■	N	N	Y	---
TDV-optimization	present work								■						■	N	N	N	Y

Legend: ■ Intrinsic to the clustering/partitioning approach. □ Optional methodological pathway. **Y** Yes. **N** No. **O** Optionally. **P** Partially. In green, properties of the approach that are consistent with the tabulation technique as described in Ellenberg (1956); in red otherwise.

[1] Heuristics include, e.g., iterative reallocation algorithms with a fast/greedy strategy. Usually, metaheuristics also use iterative reallocation algorithms but explore the search space more efficiently. [2] These approaches do not return a partition of the relevés, but biclusters. [3] This approach returns species clusters, i.e., it is an inverse analysis (see Kent and Coker 1992, p. 279). [4] With parameter Y = 0%, the procedure drops the concentration approach and ensures exclusiveness of differential species, but this was not explored by the authors. [5] The ordination phase is obtained with all species, but the fine-tuning of the division is informed by a subset of (pseudo)species. [6] The PAM phase uses a pairwise Euclidean distance matrix obtained from the ordination space. [7] As distance to medoids is calculated using the Euclidean distance matrix obtained from the ordination space, the clusters are expected to be more irregularly shaped in terms of conventional dissimilarity measures. [8] Even if the optimal number of clusters can (optionally) be determined by a subset of species, note that all species are used equally in the group formation. [9] The graph connects relevés based on similarity indices. [10] The graph is built using all species, but edges are optionally dropped if relevés have different dominant species. [11] The graph connects species based on a co-occurrence index. [12] The graph is built using all species, but, further in the procedure, some species with co-occurrence index lower than a predefined threshold can be optionally removed from the graph.

the space used to represent the phytosociological table, the tabulation process is purely combinatorial and data-driven, with *no shape imposed on the relevé groups*.

(3) Equally informative species

In vegetation classification, species occurrences are not all equally informative (Mueller-Dombois and Ellenberg 1974; Kent and Coker 1992; Schmidtlein et al. 2010). However, many approaches treat all species equally, giving them the same relevance (see, e.g., Kent and Coker 1992, pp. 279). Some methods might weigh species by their cover values, but species with the same cover are still considered equally informative. In phytosociological tabulation, all species are initially considered informative, but only differential species substantiate the final partition – i.e., only a subset of the species is truly informative for the identified partition. Kent and Coker (1992, pp. 279–280) discussed two closely linked features: the "equal emphasis of species" and the "joint classification of quadrats and species simultaneously" (i.e., biclustering of relevés and species). The main goal of phytosociological tabulation is not the biclustering of the table itself, but the partitioning of relevés informed by these biclusters. In summary, the final output of the tabulation respects a certain biclustering structure (which I will describe later), in which *species are not equally informative* (see also Dale 1988).

(4) Type of informative species

For those methods that recognize the relevance of subsets of species in the emergence of relevé groups, it is crucial to *ensure the exclusivity of these species to one or more emerging groups*, as this is a fundamental aspect of the tabulation technique. Note that several modern approaches to quantifying fidelity or indicator value use a concentration-based approach, which does not guarantee exclusivity to specific groups (see Appendix 1 for a detailed discussion of this issue). Relying on a concentration-based approach when identifying subsets of informative species may lead to different subsets from those obtained through phytosociological tabulation.

When the clustering/partitioning approach does not meet these four requirements of the phytosociological tabulation (italicized above), it risks misplacing relevés compared to the tabulation technique.

Madeira and Oliveira (2004) reviewed biclustering methods. I use their terminology to describe biclustering structures. The traditional tabulation technique aims to achieve *exclusive-row and -column* biclusters (see figure 1b in Podani and Feoli 1991; or figure 4 in Madeira and Oliveira 2004). Figure 1A depicts an *exclusive-row and -column* bicluster structure (dark grey and white areas), but *non-exhaustive-in-the-rows* (note the light grey area at the bottom, representing rows not attributed to any bicluster). The relaxation of the row's exhaustiveness is significant in vegetation science, considering the high occurrence of dominant and companion species that may span all groups under comparison. Since the ideal scenario of an *exclusive-row and -column* diagonal of blocks is rarely observed in real-world data, there is an additional relaxation of column exclusiveness. This means that only *exclusive-row* biclusters are sought, where relevés can be associated with multiple dark grey blocks, but each species is still assigned to one block only. This adjusted structure is shown in Figure 1B.

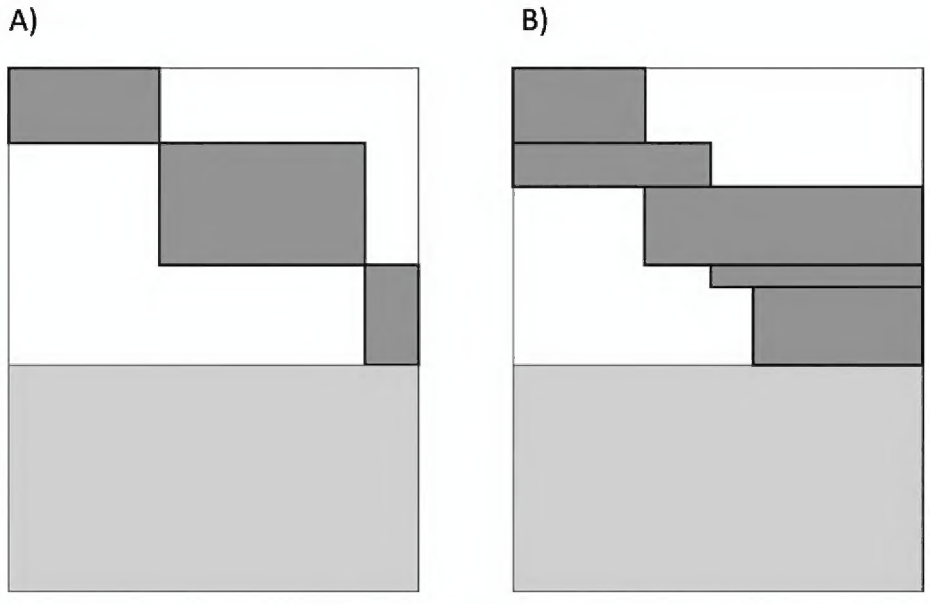

Figure 1. Diagrams of bicluster structures (species in rows, relevés in columns). Dark and light grey blocks enclose the occurrences of species, while white areas contain only the absences of species. **A)** An *exclusive-row and -column, non-exhaustive-in-the-rows* structure. **B)** A relaxation of the previous structure, where only *exclusive-rows, non-exhaustive-in-the-rows* biclusters are depicted.

In manual tabulation, mutually exclusive groups of all columns (i.e., a partition of the relevés) are adjusted to the identified bicluster blocks to reveal exclusive species patterns among the final relevé groups. Figure 2A illustrates how a partition (represented by groups I, II and III) is overlaid on the bicluster structure from Figure 1B, showing how the groups are differentiated by the presence (or presence and absence) of certain species. The block shapes and groups in Figure 2A roughly depict the final tabulation from Mueller-Dombois and Ellenberg (1974) for the *Arrhenatheretum* data set.

The bicluster structure in Figure 2B simplifies that of Figure 2A: the dark grey blocks, which enclose all presences of a species, are now aligned with groups I, II, and III, merging where appropriate into unique blocks. These blocks represent sets of differential species for a given group or combination of groups. For example, species 9 is a differential species for groups II and III because it is absent from group I. *Nota bene* 1: White areas (where only absences are recorded) ensure that a species is absent from at least one group – an essential property for differential species under the definition used here. *Nota bene* 2: In Figure 2B, all species within a block have at least one presence in each associated group. For clarity, if species 9 lacked its two presences in group II, it would instead be a differential species of group III and placed in the dark grey block below.

Recall that the bottom light grey area contains species left outside the biclusters. In phytosociological tabulation, some species absent from one or more relevé groups may remain in this area, though this can be considered subjective. In an automatic classification, however, each species is preferentially assigned to a unique block, ensuring that each block contains only species present in all its associated groups. Any later adjustments can be justified by the expert. Thus, in Figure 2B, the light grey area includes only species occurring at least once in all three groups (I, II, and III), exemplified by species 15.

Looking carefully at Figure 2B, it is possible to recognize two different kinds of absences. The first kind, represented in red, is inside the dark grey or the light grey areas, where it is possible to find absences together with occurrences of the species (remember that species associated with these areas show at least one presence in the associated groups). If present, such absences are generally seen as *stochastic absences*. The absences of the second kind (represented in green) occur inside the white areas, where there are no presences at all. As these absences allow the species to differentiate certain groups, I call them *differentiating absences*. The former can be attributed to stochastic processes that determine the absence of a species from a plot, even if such species is present in the species pool of the respective group(s). The latter can be attributed to a complete absence from the species pool of the respective group(s), e.g., due to biogeographical-related causes or strong environmental filtering.

The Total Differential Value (TDV), proposed in this work, seeks to answer the following question: "Can a numerical index be created to globally reflect the strength of

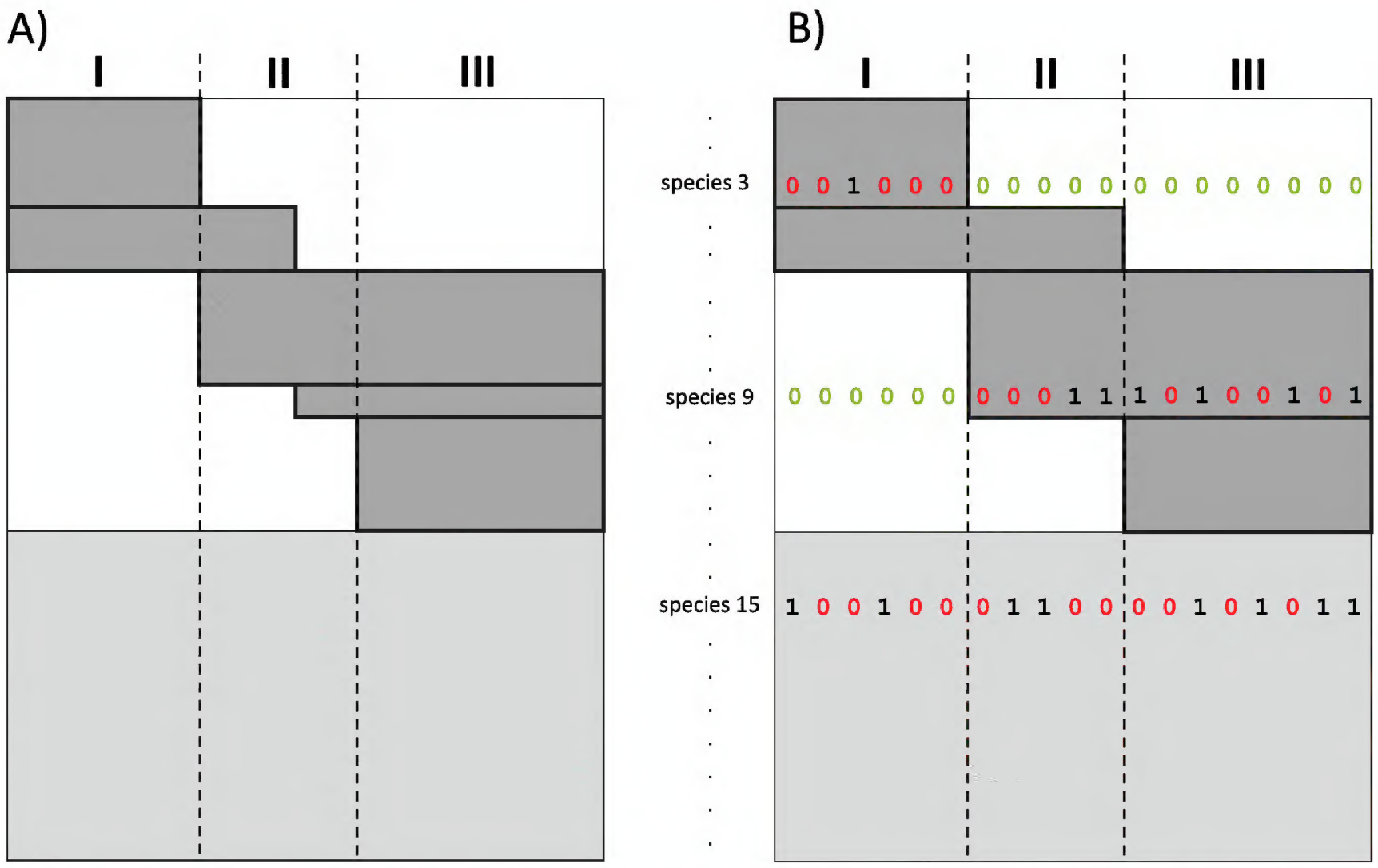

Figure 2. Diagrams of bicluster structures and partitions of the columns/relevés as sought in the phytosociological tabulation. **A)** A partition of the columns/relevés (groups I, II and III) superimposed on the biclustering structure from Figure 1B. **B)** Illustration of some hypothetical species presences (1) and absences (0) over a simplified version of the biclustering structure shown on the left. Stochastic absences are shown in red, and differentiating absences are in green (see text for further explanation).

differential species patterns in a differentiated table?" To answer this question, TDV leverages the distinction between stochastic and differentiating absences.

The total differential value

An ideal block of differential species is characterized by a high density of presences, coupled with a total absence from at least some of the remaining groups, preferably from all. An ideal tabulation would display large blocks of exclusive species, differentiating each group. However, in practice, the number of species between blocks varies, as does the density of occurrences of each species within the same block.

The usefulness of a single species in differentiating a specific group of relevés increases with (i) the frequency of the species' presences in the group, and (ii) its exclusivity to the group, meaning it is absent from the other groups as much as possible. This feature is commonly sought in species indicator power measures (e.g., Dufrêne and Legendre 1997) and clustering sharpness measures (e.g., Podani and Feoli 1991). The maximum values of these measures occur when a species is present in all sites of one group, and only absences are recorded outside that group. The Differential Value (DiffVal), presented below, exhibits this same behaviour while introducing an innovative approach to considering absences outside the groups.

Given a phytosociological table T and a k-partition P of its relevés (i.e., a classification of all relevés into k groups, with each relevé assigned to only one group), the DiffVal expresses the ability of a species, say s, to differentiate the groups in which it occurs from the other groups. The $DiffVal$ of species s, given partition P, is calculated by the following formula:

$$DiffVal_{s,P} = \frac{1}{e} \sum_{g=1}^{k} \frac{a_g}{b_g} \times \frac{c_{g'}}{d_{g'}} \qquad (1)$$

where a_g is the total number of presences of species s within group g; b_g is the total number of relevés in group g; $c_{g'}$ is the total number of differentiating absences of species s in groups other than g; $d_{g'}$ is the total number of relevés in all groups excluding group g; and e is the total number of groups in which species s occurs at least once.

DiffVal is, therefore, a summation of k summands, each one representing the partial differential value of the species s to each one of the k groups. Note that a_g/b_g is the frequency of the presences (of species s inside group g), relative to the group g size (i.e., the number of relevés in g). This is the constancy of species s to group g (Mueller-Dombois and Ellenberg 1974). The ratio will equal its maximum value, 1, if, and only if, species s occurs in all the relevés of group g. Additionally, note that $c_{g'}/d_{g'}$ is the frequency of the differentiating absences (of species s outside group g), relative to the sum of the sizes of all groups but g. That is, absences in $c_{g'}$ are counted outside the group g but only in the groups from which species s is completely absent (as

these are the relevant absences, which produce differentiation among groups). In practice $c_{g'}$ corresponds to the sum of the sizes of all groups other than g that are empty. Thus, $c_{g'}/d_{g'}$ will equal its maximum, 1, if, and only if, species s is completely absent from all groups but g. Finally, $1/e$ ensures that DiffVal is a value from 0 to 1. Note that DiffVal is simply based on counting the presences and absences of a species, and the size of the given groups (i.e., it is grounded in combinatorial-discrete mathematics) and was devised to measure the differential power of a species.

The TDV is the mean of the DiffVal outcomes for species in table T. It is a global measure (ranging from 0 to 1) of how well the k groups can be distinguished from one another using differential species. Thus, the TDV of a partition P over the phytosociological table T is given by:

$$TDV_{P,T} = \frac{1}{n}\sum_{s=1}^{n} DiffVal_{s,P} \qquad (2)$$

where n is the total number of species in table T and $DiffVal_{s,P}$ is the differential value of species s, given partition P.

Table 2 exemplifies the calculation of DiffVal and TDV on a hypothetical matrix of seven species and nine relevés partitioned into three groups. For computation purposes it is worth mentioning that when a species is not present in a group, the respective DiffVal summand is always zero (since the relative frequency a_g/b_g equals zero when $a_g = 0$), meaning that the summands only need be computed for the groups containing the species. Similarly, when the species occurs in all groups, all summands are zero (since $c_{g'}/d_{g'}$ equals zero for all summands where $c_{g'} = 0$). In such cases, performing the calculation is unnecessary, as DiffVal will also be zero. For illustrative purposes, all calculations are shown in Table 2.

Note that an interesting property emerges from the way the distinction between stochastic and differentiating absences is incorporated into DiffVal. Specifically, the distribution of differentiating absences among the groups under comparison allows for the establishment of exclusiveness degrees (i.e., species exclusive to one group, two groups, three groups, etc.), quantifying the differential power of the species and ultimately distinguishing differential from non-differential species. The latter species, which occur in all vegetation groups of the phytosociological table, do not contribute to TDV, as their respective DiffVal is zero.

Optimization of TDV

The number of possible partitions of a set with n elements is well-studied in combinatorics and corresponds to the n^{th} Bell number (Bell 1934), usually denoted B_n. For example, 10 relevés can be partitioned in B_{10} = 115,975 different ways. If we constrain the number of groups, say to k, the possible partitions with k groups (of the same n elements) correspond to the Stirling number of the second kind (Boyadzhiev 2012), usually denoted $S(n,k)$. For instance, the same 10 relevés can be partitioned into three groups in $S(10,3) = 9{,}330$ different ways, and into five groups in $S(10,5) = 42{,}525$ different ways. The "Stirling2" function from the R package *multicool* (Curran et al. 2024) implements these calculations.

The TDV metric may provide a way to assist tabulation. Specifically, if we accept that DiffVal is a good measure of the differential value of a species, and if we accept that TDV is a good measure of how well a partition is characterized in terms of the strength of the differential species patterns it contains, then optimizing TDV would allow us to find tabulations close to the ones sought in Braun-Blanquet's school.

As Bell numbers and Stirling numbers of the second kind increase rapidly, calculating TDV for all possible partitions of the relevés in a phytosociological table is often impracticable. In other words, complete enumeration is not a viable option, as is the case with other criteria (see, e.g., Podani and Feoli 1991; Peet and Roberts 2013). However, appropriate optimization techniques (e.g., local search-based metaheuristics, also known as non-hierarchical iterative reallocation algorithms) can be used to maximize TDV, as will be shown hereafter.

R package *diffval* (Monteiro-Henriques and Cerdeira 2023) includes functions dedicated to the optimization of TDV. In the following subchapters, I will illustrate TDV-optimization using variations of the hill climbing algorithm (a local search technique; see, e.g., Basseur and Goëffon 2015), implemented in function "optim_tdv_hill_climb". The function's documentation provides a detailed description of the optimization procedure. Monteiro-Henriques and Cerdeira (2023) also provide functions for a mixed-integer linear programming approach (for the two-group case) and a simulated annealing algorithm (for two or more groups). Additionally, these authors are working on a mathematical proof showing that finding a partition that maximizes TDV is NP-hard (a class of problems in computational complexity theory) for two or more groups.

Table 2. Example of DiffVal and TDV calculation on a hypothetical matrix.

relevé no.	12|345|6789	DiffVal calculation	DiffVal summands	DiffVal
group no.	11|222|3333			
species 1	11|000|0000	1/1 × [(2/2) × (7/7) + (0/3) × (4/6) + (0/4) × (3/5)] =	1/1 × [1.00 + 0.00 + 0.00] =	1.00
species 2	00|000|1011	1/1 × [(0/2) × (3/7) + (0/3) × (2/6) + (3/4) × (5/5)] =	1/1 × [0.00 + 0.00 + 0.75] =	0.75
species 3	00|101|0000	1/1 × [(0/2) × (4/7) + (2/3) × (6/6) + (0/4) × (2/5)] =	1/1 × [0.00 + 0.67 + 0.00] =	0.67
species 4	11|000|1111	1/2 × [(2/2) × (3/7) + (0/3) × (0/6) + (4/4) × (3/5)] =	1/2 × [0.43 + 0.00 + 0.60] =	0.51
species 5	01|100|0000	1/2 × [(1/2) × (4/7) + (1/3) × (4/6) + (0/4) × (0/5)] =	1/2 × [0.29 + 0.22 + 0.00] =	0.25
species 6	11|111|1101	1/3 × [(2/2) × (0/7) + (3/3) × (0/6) + (3/4) × (0/5)) =	1/3 × [0.00 + 0.00 + 0.00] =	0.00
species 7	10|101|0100	1/3 × [(1/2) × (0/7) + (2/3) × (0/6) + (1/4) × (0/5)] =	1/3 × [0.00 + 0.00 + 0.00] =	0.00
				TDV = 3.18/7 = 0.45

In the *diffval* package, users can also find the "tabulation" function, which, given a phytosociological table and a partition of its relevés, rearranges the table's rows and columns to display exclusive species at the top. The reordering follows these steps: (i) Species are first ordered by the increasing number of groups in which they occur (i.e., by exclusiveness level/degree). (ii) Within each exclusiveness level, species are further ordered lexicographically based on the groups to which they belong. (iii) Finally, within each level defined by the previous two steps (also known as shortlex order), species are reordered by the decreasing sum of their relative frequencies across the groups in which they occur.

The columns are also reordered according to the increasing order of the assigned group membership numbers. Optionally, the rearranged table can be visualized graphically (an example is provided below).

It is important to note that the tabular rearrangement performed by the "tabulation" function is based on simple combinatorial rules (primarily the shortlex order). Its effectiveness in generating a meaningful phytosociological tabulation depends entirely on the data and the given partition.

Examples

I illustrate TDV-optimization using an artificial data set (Example 1), and a real-world data set (Example 2). I contrast TDV-optimization with ten clustering/partitioning methods (see Table 3) using seven statistics (abbreviations in bold): (1) **PARTANA**, the ratio of within-cluster similarity to among-cluster similarity (Roberts 2015), (2) the average silhouette width (**ASW**, Rousseeuw 1987), (3) the minimum silhouette width (**MSW**, Lengyel et al. 2021), (4) **OptimClass 1**, the total count of faithful species across all clusters of the partition, at significance levels of 0.05, 0.01, 10^{-3} and 10^{-6} (Tichý et al. 2010), (5) the sum of significant indicator values (IndVal) assigned to the cluster where each species has its highest indicator value, at $\alpha = 0.05$ (**ssIndVal**, see Dufrêne and Legendre 1997; Roberts 2015), (6) the number of significant indicator species that contribute to ssIndVal (**nsIndic**, see Dufrêne and Legendre 1997), and (7) The average ISAMIC score (**avISAMIC**, see Roberts 2015). For both examples, I provide additional statistics such as the number of discrepant relevés in comparison to known partitions, the number of detected differential species, and, of course, the TDV.

IndVal p-values are calculated using permutations. To obtain stable estimates, ssIndval and nsIndic are averaged over 100 runs of the "indval" function from the *labdsv* package (Roberts 2023), with 9999 randomizations per run.

For all methods that required dissimilarity matrices, the Bray-Curtis index (Bray and Curtis 1957) was used, calculated using the *vegan* (Oksanen et al. 2024) and *labdsv* (Roberts 2023) packages, as it is one of the most widely used metrics in vegetation science. I focused on three groups, as the artificial data set in Example 1 follows a three-group structure, and the real-world data set in Example 2 is traditionally classified into three groups.

OptimClass 1 was calculated using Fisher's exact test for the right-tailed hypothesis (Tichý et al. 2010), ensuring that each species was counted only once, even when the test was significant for two groups. Note that these methodological choices are implemented differently in the function "optimclass" from the *optpart* package (Roberts 2020). Suppl. material 1 contains the functions used to calculate OptimClass, adapted from Roberts (2020).

Example 1: An artificial data set

Reticulated patterns are misaligned subsets of species that produce multiple, sound groupings of relevés (see also Dale 1988). Holzner et al. (1978) provided an extreme example of reticulate patterns of differential species,

Table 3. Summary of clustering/partitioning methods compared to TDV-optimization, along with their abbreviated names used hereafter.

Method	Abbreviated name (in bold) and software package	Parameters
Agglomerative hierarchical clustering	**Ward**, *cluster*[1]	BCD, Ward's method, dendrogram cut into three groups
	Flexible β, *cluster*[1]	BCD, Flexible β method, β = -0.25, dendrogram cut into three groups
Centroid partitioning (*k*-means)	***k*-means**, *stats*[2]	number of groups = 3, no. of runs/starts = 10
Modified-TWINSPAN (Roleček et al. 2009)	**Mod-TWINSPAN**, *twinspan*[3]	cut levels = 0, max. no. of indicators for division = 7, min. group size for division = 5, max. depth of levels of divisions = 6, dendrogram cut into three groups, using class heterogeneity (Roleček et al. 2009)
Partition around the medoids	**PAM**, *cluster*[1]	BCD, no. of groups = 3
Divisive analysis clustering	**DIANA**, *cluster*[1]	BCD, dendrogram cut into three groups
Isometric feature mapping and partitioning around medoids	**ISOPAM**, *isopam*[4]	BCD, no. of groups = 3
Partitioning by optimizing PARTANA ratio	**OPTPART**, *optpart*[5]	BCD, desired no. of groups = 3 (optimization parameters given in text)
Partitioning by optimizing mean silhouette width	**OPTSIL**, *optpart*[5]	BCD, desired no. of groups = 3 (optimization parameters given in text)
Reallocation of misclassified objects using the silhouette width criterion	**OPTSIL+REMOS**, R code supplied in Lengyel et al. (2021)	REMOS1 application to all OPTSIL local optima. BCD, threshold of silhouette width for misclassified objects = -0.001
Partitioning by optimizing TDV	**TDV-optimization**, *diffval*[6]	no. of groups = 3 (optimization parameters given in text)

References to software packages: [1] Maechler et al. (2023), [2] R Core Team (2024), [3] Oksanen and Hill (2025), [4] Schmidtlein et al. (2010, 2025), [5] Roberts (2020), [6] Monteiro-Henriques and Cerdeira (2023). **Parameter abbreviations:** BCD, Bray-Curtis dissimilarity.

highlighting the challenges of treating these collectively. As a solution, the authors proposed working with subsets of species to obtain appropriate clusters. Since DiffVal does not assign differential value to species present in all groups, I hypothesize that TDV-optimization, using a local search-based metaheuristic such as random restart hill climbing, can focus on different subsets of the relevé matrix. This approach may help detect reticulate patterns among partitions with higher TDV.

To test this, I created an artificial data set (107 species × 64 relevés) containing two distinct patterns of differential species distributed across three groups. The first pattern included three blocks of 16, 18, and 30 relevés, with 6, 8 and 12 differential species, respectively. The second pattern comprised three additional blocks of 14, 22, and 28 relevés, with 5, 11, and 9 differential species, respectively. These two patterns were randomly interwoven, ensuring that the blocks composing each pattern remained misaligned. The remaining 56 species were considered common-to-rare, without any associated pattern. Species presences within each block were randomly generated with varying theoretical relative frequencies, decreasing exponentially from 0.9 to around 0.1. A similar process was applied to the 56 common-to-rare species, but their presences were distributed across all 64 relevés, with frequencies decreasing exponentially from 1.0 to around 0.05.

The first pattern of differential species defined three groups of relevés, labelled 1, 2, and 3 (hereafter pattern '123'), while the second pattern defined three groups labelled a, b, and c (pattern 'abc'). Each relevé was assigned both a number and a letter based on the differential species it contained (Figure 3). Figure 3A makes pattern '123' apparent by sorting relevés according to their numerical labels, whereas Figure 3B reveals pattern 'abc' by sorting them by their letter labels. Species occurrences in each block of differential species are indicated by different colours, while the 56 common-to-rare species are shown in grey. The code to generate the artificial data set is provided in Suppl. material 2.

I submitted this data set, with relevés in random order, to the classification procedures listed in Table 3. Recall that there are $S(64,3)$ possible ways to partition the 64 relevés into three groups – an astonishing 7.6×10^{18} different solutions to this specific classification problem.

Optimization procedures (OPTPART, OPTSIL, and TDV-optimization) were run multiple times with random initial partitions, retaining local optima for further analysis. OPTPART was run 50,000 times with default settings: a maximum of 100 iterations per run and a minimum PARTANA ratio increment of 0.001 to continue iterating. OPTSIL was run 20,000 times with default values for the maximum of 100 iterations per run. TDV-optimization was run 5,000 times using the "optim_tdv_hill_climb" function (Monteiro-Henriques and Cerdeira 2023). Each run began with 1,500 iterations of stochastic hill climbing, followed by up to 30 iterations of greedy hill climbing (see Basseur and Goëffon 2015). These iteration numbers were chosen based on initial tests, which confirmed that TDV plateaued in single runs. The 500 highest-TDV partitions were retained for further analysis. The minimum group size (min_g_size) was set to the default value of 1.

For each optimization procedure, I analysed the partition with the highest observed value of the optimized criterion and searched the local optima for partitions closest to patterns '123' and 'abc'. For the OPTSIL+REMOS method, I also analysed the solution with the highest ASW and searched the local optima for partitions as close as possible to these patterns. Partition discrepancy (or distance) was measured as the minimum number of relevés that must be reassigned to obtain one partition from another (see MINDMT in Podani 1986). Hereafter, I refer to this metric as "number of

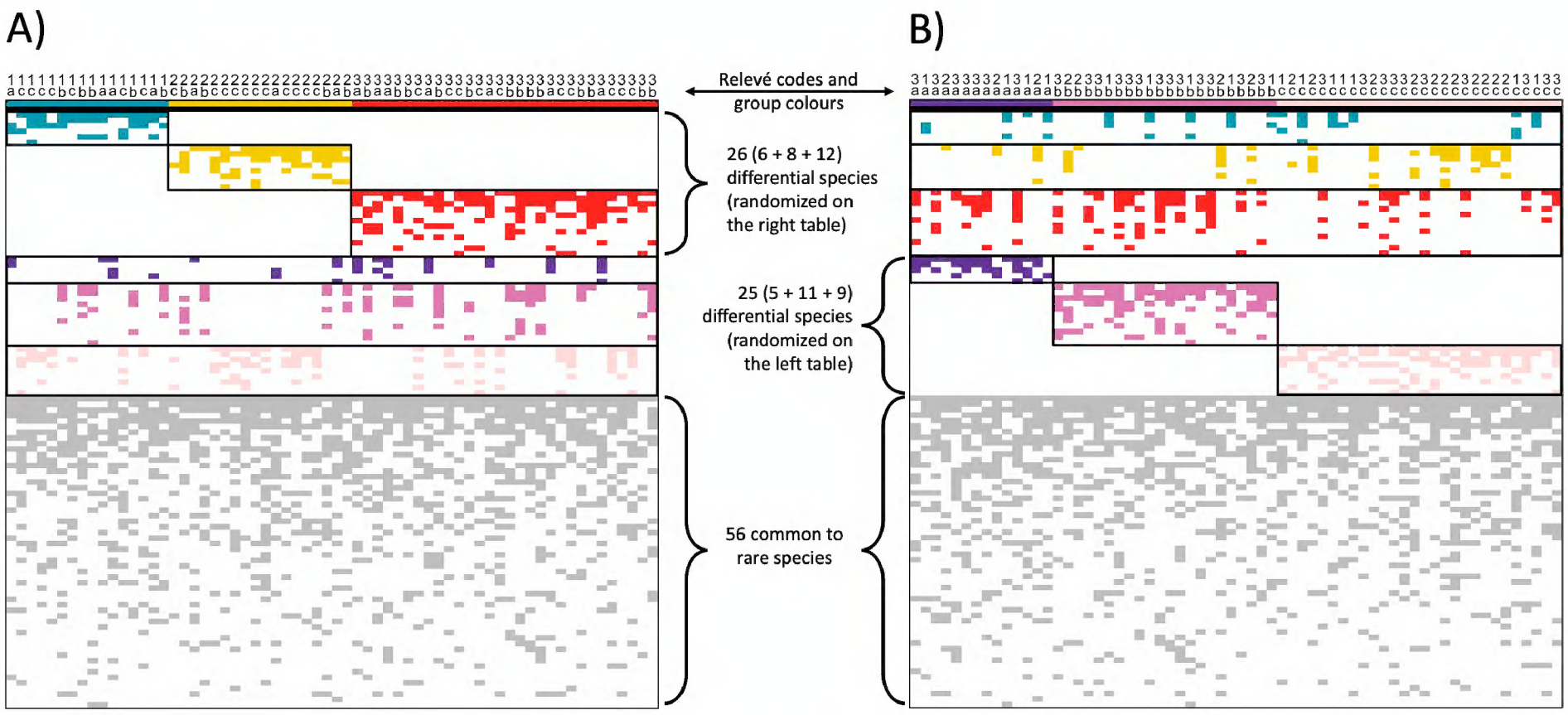
A)
B)
Relevé codes and group colours
26 (6 + 8 + 12) differential species (randomized on the right table)
25 (5 + 11 + 9) differential species (randomized on the left table)
56 common to rare species

Figure 3. Artificial data set displaying its two reticulated patterns of differential species. Species presences are represented as small rectangles (coloured or grey). **A)** Relevés are sorted by the numbers in their labels, revealing pattern '123' with 26 differential species. **B)** Relevés are sorted by the letters in their labels, revealing pattern 'abc' with 25 differential species.

discrepant relevés" (NDR) followed by the target partition (e.g., NDR'123' and NDR'abc'). Suppl. material 1 includes the function for calculating discrepant relevés between two partitions. Suppl. material 3 displays the graphical output of the "tabulation" function for each partition, sorting species by their degree of exclusiveness. Each output provides additional information on OptimClass 1 and IndVal significance, highlighting the species in patterns '123' and 'abc'. Alongside the tabulations, biplots of nonmetric multidimensional scaling (NMDS) and correspondence analysis (CA) are provided (Oksanen et al. 2024), highlighting the partition groups.

Table 4 shows the statistics obtained for the partitions returned by each method. In addition to the previously mentioned statistics, Table 4 includes:

(1) The number of significant species among the 56 common-to-rare species (**NSS-c-to-r**).
(2) The number of significant species among the 26 species composing the pattern '123' (**NSS'123'**).
(3) The NSS'123' that are exclusive to one or two groups (**NSSExc'123'**).
(4) The number of significant species among the 25 species composing the pattern 'abc' (**NSS'abc'**).
(5) The NSS'abc' that are exclusive to one or two groups (**NSSExc'abc'**).
(6) The number of discrepant relevés relative to pattern '123' (**NDR'123'**).
(7) The number of discrepant relevés relative to pattern 'abc' (**NDR'abc'**).

Significance for (1) to (5) was determined using Fisher's exact test, exactly as in OptimClass 1, with a significance level (α) of 0.05. The sum of NSS-c-to-r, NSS'123', and NSS'abc' equals OptimClass 1 (α = 0.05).

NSS-c-to-r counts how many of these 56 species showed a statistically significant concentration in at least one group. At the commonly accepted significance level of α = 0.05, an average of 2.8 species are expected to be falsely detected as concentrated in at least one group. Table 4 shows that some methods (*k*-means, ISOPAM, Flexible β, OPTPART, Ward and PAM) produced particularly high NSS-c-to-r values (8 to 14). These high false positive rates (14 to 25%) stem

Table 4. Statistics for the partitions returned by the tested methods for the Example 1 data set. The best scores in each row are in bold, and the maximum (or minimum) values are in bold italics.

	Agglomerative clustering							**OPTPART** selected partitions:			**OPTSIL** selected partitions:		**OPTSIL+ REMOS** selected partitions:		**TDV-optimization** selected partitions:		
	Ward	**Flexible β**	***k*-means**	**Mod-TWINSPAN**	**PAM**	**DIANA**	**ISOPAM**	maximum found	local optimum closest to '123'	local optimum closest to 'abc'	maximum found and closest to 'abc'	**'123' is among local optima**	maximum found and closest to '123'	**'abc' is among local optima**	maximum found	**'123' is among local optima**	**'abc' is among local optima**
Partition id:	A	B	C	D	E	F	G	H	I	J	K	'123'	L	'abc'	M	'123'	'abc'
PARTANA	1.207	1.183	**1.258**	1.189	1.187	1.173	1.258	***1.265***	1.240	1.255	1.249	1.228	1.256	1.243	1.246	1.228	1.243
ASW	0.0814	0.0714	0.1018	0.0729	0.0723	0.0496	0.1024	0.0989	0.0946	0.1010	***0.1042***	0.1036	**0.1039**	0.1014	0.0491	0.1036	0.1014
MSW (×10)	-0.965	-1.545	-0.087	-0.758	-0.925	-1.379	0.012	-0.336	-0.51	-0.461	-0.318	-0.229	0.002	0.022	-2.461	-0.229	0.022
OptimClass 1 (α = 0.05)[1]	**43** (25)	34 (24)	42 (33)	33 (27)	**43** (24)	29 (22)	**43 (32)**	***45 (35)***	41 (24)	31 (23)	29 (25)	26 (24)	39 (24)	28 (24)	5 (5)	26 (24)	28 (24)
OptimClass 1 (α = 0.01)[1]	***25*** (17)	16 (15)	18 (17)	19 (16)	**23** (15)	16 (15)	22 ***(19)***	20 ***(19)***	**23** (17)	17 (15)	16 (16)	18 **(18)**	22 **(18)**	17 (17)	0 (0)	18 **(18)**	17 (17)
OptimClass 1 (α = 10^{-3})[1]	10 (6)	12 (11)	11 (10)	11 (10)	11 (6)	10 (10)	11 (10)	9 (9)	**14** (11)	12 (11)	12 (12)	**14 (14)**	***16*** **(13)**	**14 (14)**	0 (0)	**14 (14)**	**14 (14)**
OptimClass 1 (α = 10^{-6})[1]	3 (3)	3 (3)	5 (4)	3 (3)	2 (1)	4 (4)	6 (5)	4 (4)	4 (3)	7 (6)	**8 (8)**	***9*** **(9)**	5 (4)	***9*** **(9)**	0 (0)	***9*** **(9)**	***9*** **(9)**
sslndval (α = 0.05)	10.66	10.41	10.64	9.82	***12.08***	7.94	**12.00**	9.99	10.88	9.76	9.74	9.61	11.89	10.01	1.87	9.61	10.02
nslndic (α = 0.05)	31.93	25.00	30.53	29.39	***36.53***	20.06	**36.37**	27.90	30.87	22.32	24.22	21.00	34.27	23.81	2.00	21.00	23.87
avISAMIC	0.6654	**0.6846**	0.6695	0.6717	0.6709	0.6794	0.6710	0.6636	0.6705	0.6657	0.6640	0.6771	0.6731	0.6670	***0.8797***	0.6771	0.6670
NSS-c-to-r (α = 0.05)	11 (20%)	10 (18%)	8 (14%)	4 (7%)	14 (25%)	5 (9%)	9 (16%)	10 (18%)	11 (20%)	**3 (5%)**	5 (9%)	***2*** **(4%)**	7 (12%)	**3 (5%)**	4 (7%)	***2*** **(4%)**	**3 (5%)**
NSS'123' (α = 0.05)	17 (65%)	**21 (81%)**	17 (65%)	16 (62%)	18 (69%)	19 (73%)	18 (69%)	15 (58%)	20 (77%)	6 (23%)	0 (0%)	***23*** **(88%)**	19 (73%)	1 (4%)	0 (0%)	***23*** **(88%)**	1 (4%)
NSSExc'123' (α = 0.05)	10 (38%)	**20 (77%)**	17 (65%)	13 (50%)	12 (46%)	19 (73%)	18 (69%)	14 (54%)	19 (73%)	3 (12%)	0 (0%)	***23*** **(88%)**	19 (73%)	0 (0%)	0 (0%)	***23*** **(88%)**	0 (0%)
NSS'abc' (α = 0.05)	15 (60%)	3 (12%)	17 (68%)	13 (52%)	11 (44%)	5 (20%)	16 (64%)	20 (80%)	10 (40%)	**22 (88%)**	***24*** **(96%)**	1 (4%)	13 (52%)	***24*** **(96%)**	1 (4%)	1 (4%)	***24*** **(96%)**
NSSExc'abc' (α = 0.05)	12 (48%)	1 (4%)	15 (60%)	12 (48%)	8 (32%)	2 (8%)	13 (52%)	18 (72%)	4 (16%)	**20 (80%)**	***24*** **(96%)**	1 (4%)	5 (20%)	***24*** **(96%)**	1 (4%)	1 (4%)	***24*** **(96%)**
NDR'123'	23	13	18	18	21	16	17	21	**12**	34	36	***0***	13	36	33	***0***	36
NDR'abc'	22	39	22	26	25	37	22	20	29	7	**2**	36	25	***0***	34	36	***0***
TDV	0.0515	0.0503	0.0615	0.0601	0.0433	0.0678	0.0586	0.0700	0.0591	0.0700	0.0904	**0.1010**	0.0511	**0.1024**	***0.1622***	**0.1010**	**0.1024**

[1] For OptimClass 1, I provide in parentheses the number of significant species that are exclusive to one or two groups.

from the fact that group formation was not independent of the distribution of these 56 species (see Discussion). These methods also yielded the highest OptimClass 1 values ($\alpha = 0.05$), ranging from 34 to 45. The solution for such an overabundance of significant species is usually to subjectively decrease α (Tichý et al. 2010; Lötter et al. 2013; Roberts 2015; Lengyel et al. 2021). However, this greatly reduces the test's statistical power. OptimClass 1 declines sharply for α values lower than 0.05, especially for these six methods. Values range from 9 to 14, for $\alpha = 10^{-3}$, and from 2 to 6, for $\alpha = 10^{-6}$. Recall pattern '123' consisted of 26 differential species, while pattern 'abc' contained 25. Suppl. material 4 provides an analysis of the statistical power of Fisher's exact test, adjusted to the scale of Example 1, illustrating the extent to which lowering α reduces differential/concentrated species detectability. Suppl. material 5 includes a graphical analysis of how OptimClass 1 varies with α across clustering/partitioning methods.

Despite the high false positive rates among the common-to-rare species, *k*-means and OPTPART detected 32 of the 51 differential species as significant (NSSExc'123' + NSSExc'abc'). ISOPAM found 31. These are notably high values, given that the corresponding partitions deviate substantially from patterns '123' and 'abc' and represent a mixture of some of the artificially created groups.

For partitions J and K, which are closer to partitions '123' and 'abc', the figures are necessarily different, showing lower values of NSS-c-t-r. Unsurprisingly, partitions '123' and 'abc' (found by OPTSIL, OPTSIL+REMOS and TDV-optimization) have OptimClass 1 values ($\alpha = 0.05$) that are closer to the actual number of differential species composing each pattern. These numbers reflect high statistical power, ranging from 88 to 96% (see the NSS of the respective pattern), and acceptable false positive rates, ranging from 4 to 5% (see NSS-c-to-r and the NSS of the other pattern).

It is worth mentioning that: (i) Mod-TWINSPAN and DIANA had relatively low NSS-c-to-r values; (ii) PAM and ISOPAM showed the highest values of ssIndVal and nsIndic; (iii) Flexible β and DIANA had relatively high values of avISAMIC; and (iv) *k*-means showed a high value for the PARTANA ratio.

The number of significant indicators (nsIndic) was consistently between OptimClass 1 values for $\alpha = 0.05$ and those for $\alpha = 0.01$.

For the optimization procedures, Table 4 shows that, as expected, each method performs well according to its criterion. OPTPART found the partition with the highest PARTANA ratio, OPTSIL the partition with the highest ASW, and TDV-optimization the partition with the highest TDV.

OPTSIL found partition '123' (among the set of local optima) and got extremely close to partition 'abc', with just two discrepant relevés (partition K, which also has the highest ASW). It is not possible for OPTISIL to find partition 'abc' exactly, as its ASW is lower than the ASW of partition K, and applying OPTSIL directly to partition 'abc' leads to convergence with partition K. The increase in ASW can only be explained by the noise that other (randomly distributed) species introduce to the ASW.

In contrast, OPTISIL+REMOS improved the MSW of partition K, converging to partition 'abc' (which is a local optimum of OPTISIL+REMOS). However, since the MSW is negative for partition '123', it becomes impossible for OPTSIL+REMOS to find it. Applying the REMOS algorithm directly to partition '123' changes 13 relevés from their original groups, leading to partition L, which has the highest ASW found by the REMOS procedure and is the closest to pattern '123'.

SillyPutty (Bombina et al. 2024) has recently been proposed as a novel clustering method, but it refines partitions by reallocating objects with negative silhouettes in the same way as the REMOS1 algorithm (Lengyel et al. 2021). I tested the "RandomSillyPutty" function (50,000 runs starting from random partitions), and the results are identical to those shown in Table 4 for OPTSIL+REMOS. Suppl. material 6 contains an analysis of how discrepant the local optima yielded by the optimization procedures (including SillyPutty) were from the patterns '123' and 'abc'.

TDV-optimization was the only method capable of converging to local optima that identified both patterns ('123' and 'abc'). However, the highest TDV corresponds to a partition (M) where two groups consist of single relevés. Partition M also exhibits the highest avISAMIC. Nevertheless, when TDV-optimization is applied with the parameter min_g_size = 4, partitions '123' and 'abc' are ranked first and second among the returned solutions (see Discussion).

Example 2: The *Arrhenatheretum* data set

Ellenberg (1956) used the *Arrhenatheretum* data set to meticulously illustrate the tabulation technique. That illustration was later translated into English by Mueller-Dombois and Ellenberg (1974) and has since become a classical data set for testing the performance of methodologies aimed at reproducing the technique (e.g., van der Maarel et al. 1978; van der Maarel 1979; Gauch 1982; Kuhn 1983; Popma et al. 1983; Wildi 1989; Bruelheide and Flintrop 1994). Table 5 lists some of these works and their performance compared to Ellenberg's partition into three groups (hereafter EllPar). The groups found by Ellenberg, as presented in Mueller-Dombois and Ellenberg (1974), are not necessarily the ground truth, but is the best solution found by the author that complies with tabulation requirements. In previous works, different authors tested various aspects of the original data set (e.g., cover vs. presence/absence), and it is not always clear whether the complete data set was used or only the portion that was analysed by Ellenberg (1956). Therefore, for a thorough comparison, I tested the performance of the clustering/partitioning methods listed in Table 3, calculating the same statistics used in Example 1. I used the original data set, trimmed to species with constancy between 10 and 60% as done by Ellenberg, and converted values to presence-absence.

For the *Arrhenatheretum* data set, OPTPART was run 5,000 times with default values for the maximum number of iterations per run (100) and for the minimum

Table 5. Research works aiming to classify the *Arrhenatheretum* data set and their performance compared to the three-group partition proposed by Ellenberg (1956).

Reference	Method	No. of discrepant relevés compared to Ellenberg's partition
van der Maarel et al. (1978)	TABORD	7
van der Maarel (1979)	Agglomerative clustering (Euclidean distance, Ward's method)	3
Gauch (1982)	TWINSPAN	7
Gauch (1982)	COMPCLUS (non-hierarchical composite clustering)	2
Kuhn (1983)[1]	VEGTAB, using the minimum spanning tree[1]	6
Popma et al. (1983)	FLEXCLUS, modified single linkage clustering + centroid clustering	8
Wildi (1989)	Agglomerative clustering (complete linkage)	3

[1]This approach does not return a partition of the relevés but a sequence to sort the relevés in the table. The sequence was divided manually into groups by the author and considering the groups given by Ellenberg, for comparison.

increment in the PARTANA ratio required to continue iterating (0.001). OPTSIL was run 5,000 times with default values for the maximum number of iterations per run (100). TDV-optimization was run 500 times using the function "optim_tdv_hill_climb". Each run began with 300 iterations of stochastic hill climbing, followed by up to 30 iterations of greedy hill climbing. These iteration numbers were chosen based on initial tests, which confirmed that TDV plateaued in individual runs. The 100 highest-TDV partitions were retained for subsequent analysis. The minimum group size (min_g_size) was set to the default value of 1.

OPTPART consistently converged to a two-group solution with a PARTANA ratio of 2.183. In the OPTPART algorithm, "no minimum cluster size is enforced" (Roberts 2020). The fact that it converged into two groups is "an indication that a partition with a fewer number of clusters achieves a better within/among similarity ratio than does a larger number" (Roberts 2020). This is consistent with the idea expressed by van der Maarel et al. (1978) and van der Maarel (1979) that a two-group partition could be appropriate to classify the *Arrhenatheretum* data set. A solution with a higher PARTANA ratio may exist for a three-group partition, but as no other partitions were returned by OPTPART (not even as local optima), this method is not included in the upcoming comparisons.

Table 6 shows the statistics calculated for the partitions of the *Arrhenatheretum* data set, as obtained by each method. The specific statistics in this example are:

(1) The number of species ***not*** considered differential by Ellenberg, but that are exclusive to one or two groups (**nEll-not-diff-but-excl**).
(2) The number of species considered differential by Ellenberg that are exclusive to one or two groups (**nEll-diff-found**).
(3) The number of discrepant relevés, considering Ellenberg's partition (**NDR'EllPar'**).

Table 6. Statistics for the partitions returned by the tested methods for the *Arrhenatheretum* data set. The best scores in each row are in bold, and the maximum (or minimum) values are in bold italics.

	Agglomerative clustering							**OPTSIL** selected partitions:		**OPTSIL+ REMOS** selected partitions:		**TDV-optimization** selected partitions:		
Partition id:	a	b	c	d	e	f	g	h	c	i	b	j	k	EllPar
	Ward	**Flexible β**	***k*-means**	**Mod-TWINSPAN**	**PAM**	**DIANA**	**ISOPAM**	maximum found	local optimum closest to EllPar	maximum found	local optimum closest to EllPar	maximum found	local optimum closest to EllPar	**Ellenberg's partition**
PARTANA	1.693	1.978	1.750	**2.015**	1.724	1.819	1.709	1.978	1.750	1.936	1.978	***2.094***	1.854	1.803
ASW	0.1556	0.1695	0.1824	0.1489	0.1679	0.1883	0.1644	***0.2280***	0.1824	**0.2079**	0.1695	0.1622	0.1697	0.1627
MSW (×10)	-1.302	0.332	-0.346	-0.637	-0.641	0.157	-0.658	-0.728	-0.346	0.020	0.332	-1.435	-1.299	-1.354
OptimClass 1 (α = 0.05)[1]	21 (18)	20 (19)	**23** (19)	21 (19)	22 (18)	***25*** (19)	***25 (21)***	15 (14)	**23** (19)	22 **(20)**	20 (19)	21 ***(21)***	22 ***(21)***	**23 *(21)***
OptimClass 1 (α = 0.01)[1]	12 (10)	10 (10)	13 (11)	12 (12)	**14** (12)	***17 (14)***	**14 (13)**	9 (9)	13 (11)	12 (12)	10 (10)	11 (11)	11 (11)	12 (12)
OptimClass 1 (α = 10^{-3})[1]	2 (2)	3 (3)	2 (2)	3 (3)	2 (2)	3 (3)	2 (2)	***5 (5)***	2 (2)	**4 (4)**	3 (3)	**4 (4)**	3 (3)	**4 (4)**
ssIndval (α = 0.05)	10.55	5.67	10.16	6.04	9.68	***11.61***	**10.81**	6.97	10.16	8.59	5.69	5.12	8.07	8.64
nsIndic (α = 0.05)	**20.82**	8.00	19.00	9.00	17.83	***21.83***	20.34	9.00	19.00	13.43	8.04	7.11	13.00	15.00
avISAMIC	0.6006	0.5997	0.6047	0.5832	0.5940	0.6150	0.5939	**0.6429**	0.6047	0.6090	0.5997	***0.7071***	0.5808	0.5824
nEll-not-diff-but-excl	9 (41%)	8 (36%)	12 (55%)	9 (41%)	**6 (27%)**	9 (41%)	***5 (23%)***	11 (50%)	12 (55%)	8 (36%)	8 (36%)	15 (68%)	***6 (27%)***	***5 (23%)***
nEll-diff-found	23 (88%)	**25 (96%)**	23 (88%)	24 (92%)	24 (92%)	24 (92%)	***26 (100%)***	22 (85%)	23 (88%)	24 (92%)	**25 (96%)**	***26 (100%)***	***26 (100%)***	***26 (100%)***
NDR'EllPar'	8	3	7	4	8	5	8	9	7	7	3	5	**1**	***0***
TDV	0.1821	0.2409	0.1909	**0.2471**	0.1838	0.2051	0.1854	0.2380	0.1909	0.2330	0.2409	***0.2511***	0.2285	0.2127

[1] For OptimClass 1, I provide in parentheses the number of significant species that are exclusive to one or two groups.

Suppl. material 7 displays the graphical output of the "tabulation" function for each partition. Each output provides additional information on OptimClass 1 and IndVal significance, highlighting the species considered differential by Ellenberg (1956).

OPTSIL found the solution with the highest ASW, followed by OPTSIL+REMOS. TDV-optimization yielded the highest values of TDV and avISAMIC. The second highest value of TDV was obtained by Mod-TWINSPAN. TDV-optimization also found the solution with the highest PARTANA ratio, again followed by Mod-TWINSPAN (but recall that OPTPART was not included in the comparison).

Ellenberg (1956) highlighted 26 differential species in the data set. OptimClass 1 for $\alpha = 0.05$ presented values close to the Ellenberg figure. As observed in Example 1, lowering α involves a sharp loss of statistical power and erratic behaviour in identifying to the *best*-performing method.

In general, all methods identified a relatively high proportion of the differential species defined by Ellenberg (nEll-diff-found), with Ward, *k*-means and OPTSIL performing relatively worse.

The local optimum closest to EllPar found by the OPTSIL algorithm is the same partition produced by *k*-means (partition c). Partition c has the highest number of species exclusive to one or two groups (nEll-not-diff-but-excl + nEll-diff-found), but the respective TDV is relatively low. Recall that DiffVal weighs the constancy in groups and the degree of exclusiveness. Partition c has a relatively low number of species exclusive to just one group.

The local optimum closest to EllPar found by the OPTSIL+REMOS procedure is the same partition produced by Flexible β (partition b), differing from EllPar by only three discrepant relevés.

ISOPAM produced an interesting partition that, despite having eight discrepant relevés (NDR'EllPar'), closely resembled EllPar in terms of differential species. The *Bromus erectus* group remains the same as in EllPar, with the difference lying in how the remaining relevés are split. The tabulation shown in Suppl. material 7: figure S7 reveals that ISOPAM's split designates *Pastinaca sativa* as a differential species for one group and *Silaus pratensis* for the other, though both exhibit low constancy within their respective groups. In contrast, EllPar (Suppl. material 7: figure S14) identifies *Deschampsia caespitosa*, *Angelica sylvestris*, *Carex acutiformis*, *Filipendula ulmaria*, and *Pimpinella magna* as differentials for one group, while the other group lacks its own differential species. The differential species block is moderately dense, but every relevé within it contains at least one species from the block.

The statistic nEll-not-diff-but-excl counts the number of species not considered differential by Ellenberg but found to be exclusive to one or two groups (and thus are potential differentials if the partition is to be accepted). High values of this measure indicate that species considered pervasive in Ellenberg's tabulation contribute to differentiate the groups of the partition under analysis. The baseline for this measure is not 0, as Ellenberg did not highlight as differential some species that were exclusive to two groups. As a result, nEll-not-diff-but-excl for Ellenberg's own partition is 5. While the decision that was made not to consider these species as differential may have been subjective (or based on expert knowledge), all had a low number of presences in the matrix: four for *Lotus corniculatus* and *Galium boreale*, and three for *Silene inflata*, *Silaus pratensis* and *Pastinaca sativa*. The automatic tabulation performed by the "tabulation" function, which is primarily based on the shortlex order, assign these species to the respective differential block (see Suppl. material 7).

Figure 4 presents the output of the "tabulation" function for partition k, which ranks ninth among the solutions with the highest TDV and is the closest to EllPar (NDR'EllPar' = 1). The only discrepant relevé is R5. This change is understandable, as *Holcus lanatus* becomes an exclusive species of a single group without the loss of any other differential species or changes in their degree of exclusiveness. The "tabulation" function outputs for the top ten solutions with the highest TDV are provided in Suppl. material 8.

Suppl. material 8: figure S1 illustrates the solution with the highest TDV (0.2511). For the most part, the top eight solutions were largely similar, generally isolating one or two relevés while maintaining two primary groups: the *Bromus erectus* group and the *Cirsium oleraceum* + *Geum rivale* group. In some cases, the isolated relevés were merely a subset of the *Bromus erectus* group, suggesting that a two-group partition may be appropriate (see Suppl. material 8: figures S2, S4–S8). The TDV-optimization result for two groups (using the same parametrization) is provided in Suppl. material 9.

Several of the top solutions from TDV-optimization (see Suppl. material 8) closely resembled partitions obtained by other methods. The first TDV-optimization solution (partition j) differed by only one relevé (R16) to partition d (Mod-TWINSPAN) and by two relevés (R16 and R12) to partition b (Flexible β and the OPTSIL+REMOS local optimum). Partition i (OPTSIL+REMOS maximum) had a single discrepant relevé (R16) compared to the third TDV-optimization solution. Partition h (OPTSIL maximum) was identical to the eighth TDV-optimization solution.

Discussion

Example 1

Reticulate patterns are not uncommon in vegetation data, and ordination techniques are typically used to explore them (Mucina and van der Maarel 1989). These patterns often arise autonomously, reflecting narrow-scale phenomena (like variations in soil moisture or nutrient content) and broad-scale phenomena (like latitudinal/altitudinal temperature gradients or biogeographic species turnover).

Example 1 demonstrated that reticulate patterns pose challenges to most clustering and partitioning methods. This is methodologically understandable: when some relevés contain only one or a few differential species per pattern, the signal may be masked by other species combinations and remain undetected by conventional dissimilarity measures.

A key sentence in Mueller-Dombois and Ellenberg (1974, p. 187) underscores this issue:

> "*The idea of diagnostic-floristic similarity applies to the number of differential species per group that are present in a relevé. It thus differs from the idea of total floristic similarity, as evaluated for standard floristic similarity relationships.*"

Clustering and partitioning strategies that treat all species as equally informative (see Table 1) risk interferences from: (i) the distribution of hypothetical differential species, (ii) the distribution of occurrences of other non-differential species, and (iii) the distribution of the cover of other non-differential species (if cover values are considered). The effectiveness of these strategies in identifying differential species patterns depends on the alignment of the first tenet with the last two. Since such alignment is not guaranteed in vegetation data, and fluctuates fortuitously, some methods can behave erratically from one case to another. When the methods incorporate species cover, different clusters may emerge by changing the weight given to cover vs. to presence-absence. This phenomenon is precisely illustrated in the works of van der Maarel (1979) and Wildi (1989) (see also Wilson 2012).

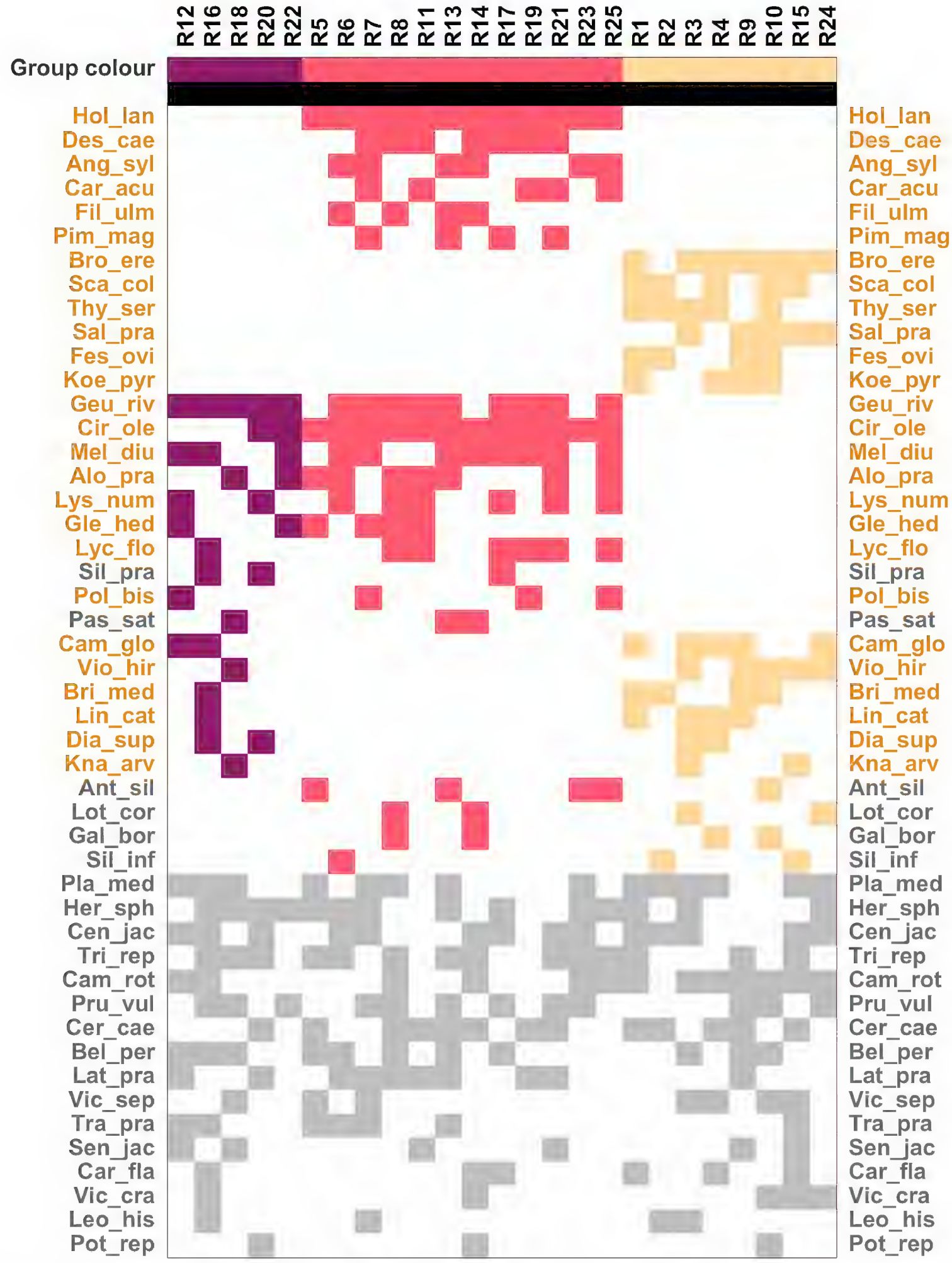

Figure 4. Visualization of the exclusive species for each group (or combination of two groups) in partition k, as generated by the "tabulation" function. Relevé and species names have been added, with Ellenberg's differential species highlighted in orange. See Ellenberg (1956) or Mueller-Dombois and Ellenberg (1974) for species names.

TDV-optimization was run with the minimum group size parameter set to 1, allowing the formation of groups with isolated relevés. The treatment of isolated objects, however, varies between methods. Some methods, like Ward and *k*-means, tend to produce relatively balanced groups and avoid isolating individual objects. In contrast, the β parameter in Flexible β offers a continuous control that balances the chaining effect of single linkage (which tends to isolate objects when cutting the dendrogram) and the more compact, balanced groups produced by complete linkage. Some metrics, such as the silhouette width, are not defined for isolated objects. The solution with the highest TDV (0.1622, Table 4) comprised two isolated relevés, set against the rest of the table (see Suppl. material 3: figure S15).

Two solutions among the local optima, in positions 196 and 197 (Suppl. material 6: figure S6), captured the artificial differential species patterns. These solutions were easy to identify, as they were the first to exclude isolated relevés. In real-world data, isolated relevés may appear either by chance or because they correspond to genuinely distinct vegetation types (this relates to the concept and treatment of outliers, which is discussed later). Since the artificial data set was specifically designed to have three clusters, the isolated relevés represent spurious patterns. By setting a larger minimum group size, solutions with isolated relevés are excluded from the search, a process akin to the effect of the β parameter in Flexible β. In this way, the remainder of the solution space is explored, potentially raising partitions that capture the artificial differential species patterns to higher positions in the solution set. I tested this by using minimum group sizes of 2, 3 and 4. The first test produced similar results, but subsequent tests with a minimum group size of 3 moved those partitions to positions 30 and 175, while setting the minimum group size to 4 placed them in the first and second positions in the solution set (Suppl. material 6: figure S7). Note that the smallest group in pattern '123' contains 16 relevés (and 14 in pattern 'abc').

The ability of TDV-optimization to find patterns is explained by its design. As DiffVal equals zero for species that pervade all groups of a given partition, TDV-optimization ignores such species in each iteration, focusing solely on the subset of the matrix where differential species may occur. The idea that it could be useful to focus on subsets of a phytosociological data set was also expressed by Wildi (1989) and Bruelheide and Flintrop (1994). If reticulate patterns of differential species are present in the data set, multiple runs of TDV-optimization, starting from random partitions, have a good chance of finding them among the local optima.

Tichý et al. (2010) treat the *p*-value of Fisher's exact test as a measure of positive fidelity and use a subjectively defined threshold (*P*) to determine faithful species. The commonly used thresholds are much lower than the typically accepted significance levels (α), which has unfavourable consequences for the test's ability to detect faithful species. Lowering α (or the *P* threshold) increases the false negative rate (usually denoted β, but not to be confused with the β in the Flexible β), thus considerably decreasing the test's statistical power (1 - β). I illustrate this in Suppl. material 4. As an example, adapted to Example 1 group sizes, consider a species with 6 presences in a group with 21 relevés (≈ 29% constancy) that is absent from the remaining 43 relevés (Fisher's exact test *p*-value = 7.24×10^{-4}). Lowering α from 0.05 to 10^{-3} decreases the statistical power (i.e., the ability to detect a species with 'true' constancy of ≈ 29%) from ≈ 96% to ≈ 58%.

The use of indices such as OptimClass (Tichý et al. 2010) should be avoided when the assessed partitions are derived from the species composition. This violates the independence assumption between the relevé groups and the species data, leading to a greater number of indicator species being identified than would be expected by chance (De Cáceres and Legendre 2009; De Cáceres et al. 2010). Milligan (1996, pp. 366–367) is highly critical of the use of hypothesis testing in this setting, noting that significant results can be obtained even from random noise. This is evident in Example 1 (see also Suppl. material 5). A strong decrease in OptimClass 1, as the α threshold departs from the usual 0.05, may reflect the spurious nature and weak concentration of such species, stemming from random fluctuations amplified by the clustering algorithms.

Example 2

One of the top partitions obtained using TDV-optimization (Figure 4), a local maximum, has one discrepant relevé compared to Ellenberg's (1956) partition. This is a relevant result, as similar attempts found in the literature have consistently shown poorer performances (see Table 5), and currently available methods do not come as close to Ellenberg's partition (see Table 6).

TDV-optimization of the *Arrhenatheretum* data set showed that presence-absence data alone was sufficient to generate a relevé partition that closely mirrors the original classification given by Mueller-Dombois and Ellenberg (1974). Therefore, the idea that presence-absence data were insufficient to allow classification of that dataset (see, e.g., van der Maarel 1979; Wildi 1989) appears flawed. This is not surprising, as it has long been known that the vegetation units in the Braun-Blanquet school are based solely on the presence of diagnostically relevant species, while species cover is used for the physiognomic characterization of communities (Tüxen and Ellenberg 1936).

TDV-optimization versus other approaches

DiffVal and TDV were designed to satisfy traditional tabulation requirements. For this reason, TDV-optimization is conceptually distinct from previously published clustering/partitioning approaches (recall Table 1). (1) Since it does not incorporate a hierarchical structure, it differs from hierarchical clustering methods (e.g., Sokal and Sneath 1963; Ward 1963; DIANA in Kaufman and

Rousseeuw 1990). (2) As it does not rely on geometric features in the search process (such as distances to centroids or medoids, or midpoints of axes), it differs from popular non-hierarchical partitioning techniques (e.g., MacQueen 1967; PAM in Kaufman and Rousseeuw 1990). (3) Frameworks that combine the methods mentioned in points (1) and (2) are therefore fundamentally different from TDV-optimization (e.g., van der Maarel et al. 1978; Hill 1979; Belbin 1987).

Most of the approaches listed in Table 1 consider species equally informative. Therefore TDV-optimization is also distinct from the methods described by Gauch (1980), Podani and Feoli (1991), Schmidtlein et al. (2010), Thiergart et al. (2014), Roberts (2015), and Lengyel et al. (2021).

Conceptually, TDV-optimization is particularly similar to biclustering approaches (like Češka and Roemer 1971; Podani and Feoli 1991; Bruelheide and Flintrop 1994; Bruelheide 2000). However, of these four approaches, only the method of Podani and Feoli (1991) returns a partition of the relevés. Their biclustering method optimizes the chi-squared statistic (using a random restart hill climbing metaheuristic). The chi-squared statistic works with presence-absence data, aiming to form blocks filled with species presences, usually represented on the diagonal of the tabulation, and blocks filled with absences outside the diagonal. According to the terminology of Madeira and Oliveira (2004), Podani and Feoli's optimization corresponds to a *non-overlapping-with-checkerboard* bicluster structure, where the maximum value of the chi-squared statistic is achieved when an *exclusive-row and -column, exhaustive-in-the-rows* bicluster structure is found. Since these bicluster structures are *exhaustive-in-the-rows*, the method compels all species into those structures, assuming they were equally informative for the partitioning of the relevés. This is not well aligned with the tabulation technique, which is clearly *non-exhaustive-in-the-rows*.

The algorithm proposed by Češka and Roemer (1971) addresses one species (or one relevé) at a time, using predefined thresholds for bicluster membership. In contrast, TDV-optimization evaluates the entire set of differential blocks simultaneously in each iteration, for each partition. Using the parameter Y = 0% in Češka and Roemer's approach would ensure the exclusivity of species to the forming biclusters, conforming to an *exclusive-rows, non-exhaustive-in-the-rows* structure. However, this approach does not seem to have been explored by the authors. This parametrization is the closest known approach to TDV-optimization (see Table 1). Nevertheless, it remains as a true biclustering method and does not return a partition of the relevés. Additionally, its implementation is heuristic rather than metaheuristic.

The ESPRESSO algorithm, proposed by Bruelheide and Flintrop (1994), is similar to the approach of Češka and Roemer (1971) in its goal of identifying blocks of co-occurring differential species. However, its reliance on a single threshold to determine the density of species presences within the emerging blocks introduces undesirable effects on block formation, particularly when compared to the tabulation technique. The threshold definition leads to a de-emphasis of low densities of presences outside the forming blocks, making it distant from the differential species concept. Specifically, a relevé containing only a few strong differential species might be excluded from block formation because it would lower the block's density. This is evident in table 8 of Bruelheide and Flintrop (1994), where species such as *Bromus erectus*, *Salvia pratensis*, *Scabiosa columbaria* and *Thymus serpyllum* in relevés 2 and 24 are assigned a block number of 0, rather than being placed in block 3, where they would become exclusive species of that block. The key difference between the ESPRESSO algorithm and traditional tabulation lies in the fact that common differential species blocks are composed of species with varying constancies, rather than blocks densely filled with presences. According to the terminology of Madeira and Oliveira (2004), the ESPRESSO approach corresponds to a flexible bicluster structure termed *non-overlapping, non-exclusive* biclusters. The COCKTAIL algorithm (Bruelheide 2000) extracts one bicluster at a time from the data matrix using a fidelity measure. However, both the density threshold in ESPRESSO and the u fidelity measure in COCKTAIL do not ensure the exclusivity of species to groups, meaning that species groups in biclusters are not necessarily exclusive to particular relevé groups. Since the COCKTAIL method was designed for large matrices, this issue is less significant. While both methods can highlight meaningfull features in the data, neither ESPRESSO nor COCKTAIL directly returns a partition of the relevés. This must be done a posteriori by examining combinations of species-relevé groups.

TDV-optimization, based on DiffVal, searches for patterns of differential species in a manner similar to Ellenberg (1956). By maximizing the mean of individual differential values, the optimization selects patterns that exhibit (i) a higher degree of species exclusiveness among the forming groups, (ii) denser blocks (i.e., containing numerous presences), and (iii) taller blocks (i.e., with more species per block). This aligns with the *exclusive-rows, non-exhaustive-in-the-rows* bicluster structure, which can readily be highlighted through a simple combinatorial reordering of the species. Furthermore, the search is conducted across the entire table and simultaneously for the predefined number of groups.

Validation measures

Brun et al. (2007) distinguish three types of validity measures: internal, relative and external. Internal indices examine the properties of clusters, such as cohesion, separation, connectivity, density, roundness, or evenness. Relative indices assess the reliability of a clustering/partitioning method applied to specific data, focusing on aspects such as stability, robustness, optimality, or reproducibility/replication. Both internal and relative indices do not require additional information about the data. In contrast, external validation requires supplementary information, such as a partition representing the ground truth.

Internal indices are probably the most commonly used (Arbelaitz et al. 2013). Within the vegetation science framework, Roberts (2015) distinguishes between two types: species-based and community-based. Examples of the former include OptimClass 1, ssIndVal, nsIndic, avISAMIC, while PARTANA ratio and ASW represent the latter. Internal evaluators have been used to compare vegetation classification methods in studies such as Aho et al. (2008), Schmidtlein et al. (2010), Lötter et al. (2013), Roberts (2015) and Lengyel et al. (2021). Pakgohar et al. (2024) tested the performance of 27 internal evaluators on noisy binary data, identifying some (including ASW and OptimClass) as the most effective. However, the reliance on OptimClass should be critically reassessed (see the discussion on Example 1).

Relative evaluators are used less often. Pillar (1999) proposed a method for assessing partition stability (see also Illyés et al. 2007; Botta-Dukát 2008).

Considering external validity indices, a recent example is Lengyel et al. (2021), who used simulated data. However, external evaluators, such as a ground truth partition, are not available in real-world vegetation analysis. Instead, researchers rely on indirect external validation using proxies for ground truth, such as environmental variables (Orlóci 1975, p. 127; Feoli and Ganis 2019, 2021). A floristic pattern without an associated (paleo)ecological interpretation lacks epistemological value (Milligan 1996, p. 365; see also Willner 2006).

A major finding of Brun et al. (2007) must be emphasized: internal and relative indices, and even some external ones, show no correlation with misclassification errors. This highlights how these indices can be disconnected from the relationship between partitions and reality, with significant epistemological implications. Furthermore, there is no general agreement on what constitutes true clusters (Hennig 2015). Hennig argues that the epistemological value of clusters lies not in their uniqueness but in the transparent and open communication of the underlying cluster concept. Brun et al. (2007) conclude that, in relation to clustering being scientifically constitutive, the historical evolution of validity indices may be considered premature. I agree with these views.

Under purely exploratory analysis, multiple clustering/partitioning methods can be applied to gain potential insights. The performance of these methods can be evaluated in relation to the vegetation phenomenon (e.g., some methods prioritize dominant species, while others highlight rare species). External validation is often compelling, particularly when groups accord with environmental gradients or reveal biogeographical patterns. Expert validation is also valuable in assessing whether the resulting groups agree with existing knowledge, relate to established vegetation classifications, or serve a practical purpose (see, e.g., Lötter et al. 2013).

I assume that vegetation scientists may also be interested in specific methods for partitioning relevés based on differential species (see Dale 1995), as these closely resemble traditional phytosociological tabulation outputs. This approach imposes precise restrictions, significantly limiting the available alternatives. Unlike exploratory analyses, it requires an applied perspective (see Milligan 1996, p. 355; Hennig 2015). The researcher must identify or develop a method that aligns with the concept of differential species and their role in phytosociological tabulation. TDV is one possibility. By distinguishing stochastic from differential absences, TDV applies the concept of exclusiveness degree to groups, ultimately distinguishing differential from non-differential species (i.e., those occurring in all groups). I have not found a similar method in the publications I consulted (see the reference list in Suppl. material 10).

Under such an applied approach and optimization framework, the optimized criterion also serves as an internal evaluator (this is self-evident, as there would otherwise be no justification for its optimization). Other internal indices may highlight characteristics that are not necessarily aligned with the concept of differential species. Species-based evaluators (other than the one being optimized) may provide insight but are unlikely to outperform the optimized criterion, especially if their conceptual basis diverges from Ellenberg's concept. As expected, indirect external validation (using proxies for ground truth and expert knowledge) functions in the same way as in exploratory analyses and should play a crucial role in validating solutions.

Outliers: Extreme values, data errors or distinct vegetation types?

In heterogeneous data sets, the highest-TDV partitions may include small groups of relevés or even isolated relevés containing some exclusive species, which some authors may consider outliers. Such isolated relevés may arise from stochastic processes, as demonstrated in the analysis of the artificial data set in this study. However, they may also represent under-sampled vegetation units with a small set of differential species.

Given that DiffVal and TDV were specifically designed to identify patterns of exclusive species, the isolation of such relevés is not only expected but also desirable. The decision to (i) impose a minimum group size (thus forcing these extreme relevés to cluster with others), (ii) exclude them from the analysis (e.g., if they contain errors or exhibit distinct physiognomies), or (iii) accept the isolation as meaningful (e.g., as distinct vegetation types), possibly increasing the total number of groups to accommodate them, must always be guided by expert knowledge, which is not embedded in the data table. Consequently, TDV-optimization cannot be expected to automatically resolve cases involving isolated relevés, if present.

Final remarks

TDV-optimization provides a means of identifying partitions that are likely to be relevant for vegetation classification. However, the researcher must still examine the solutions, possibly assessing the distinctiveness/separability of

groups in the geographical space or environmental hyperspace (Orlóci 1975; Schmidtlein et al. 2010). Indices such as those proposed by Cerdeira et al. (2012) or Feoli and Ganis (2019) may assist in this evaluation.

By design, TDV-optimization should be robust to spatial autocorrelation, as the species pools associated with a given biogeographical or environmental unit are not expected to change with the inclusion of geographically close relevés. However, a substantial number of geographically proximal relevés sharing a significant number of exclusive species may ultimately form a group. This, however, is a desirable outcome, as it reflects a genuine floristic pattern that should be detected and analysed.

It is also worth noting that, when applied to two relevés, the DiffVal reduces to the Jaccard distance, i.e., the complement of the Jaccard index (Jaccard 1901; see Murphy 1996, for an earlier reference). For an elegant generalization of the Jaccard index, see Koch (1957).

Since the inception of vegetation science, different schools of thought have emerged across regions, closely tied to their respective regional ecological contexts. By identifying patterns of differential species in vegetation data, TDV-optimization is expected to be particularly useful in floristically rich regions with high geographical turnover and/or strong environmental filtering, such as the Mediterranean region. In areas where floristic variation is primarily reflected in differences in species cover-abundance or fluctuations in species relative frequency, TDV-optimization may have limited utility, unless broad environmental gradients are analysed.

TDV-optimization is not an instantaneous method yielding a single answer; rather, it is a tool for exploring vegetation data in search of meaningful and useful classifications, potentially revealing different patterns within the same data (see Dale 1988; Schmidtlein et al. 2010), that may be harder to spot with other methodologies. These patterns can tell specific stories about the same data set.

Generality of the method

The partitioning method presented in this article, along with all its specific details, emerged from within a vegetation science conundrum, generally referred to as *vegetation classification* (see Lötter et al. 2013). While it is evident that this method could have applications in other scientific fields focused on partitioning binary data (e.g., defining subpopulations through exclusive genes blocks), I strongly believe that it was essential to present the method within the framework of vegetation science, where it has been sought for so long. This ensures that readers are familiar with the foundational research and key researchers whose work has progressively built an impressive body of knowledge, upon which this study is based.

Acknowledgments

I dedicate this work to three professors who were instrumental in shaping this study. First, Dieter Mueller-Dombois (posthumously), whom I was fortunate to meet in 2005 during the annual meeting of the International Association for Vegetation Science. Dieter left a lasting impression on me, and his seminal work, alongside that of Heinz Ellenberg, inspired this research many years ago. Second, José Carlos Costa, whose profound knowledge of flora and chorology guided me in the practice of the tabulation technique and the intricate description of vegetation communities. He patiently answered the myriad questions I raised throughout my PhD, enabling me to gain a deep understanding of the underlying principles of the method. Finally, Jorge Orestes Cerdeira, a brilliant and sharp mathematical mind with an exceptional commitment to interdisciplinary dialogue. His great patience and deeply insightful teachings on discrete mathematics – particularly graph theory, combinatorics, logic, optimization, and computer science – equipped me with the tools needed to develop DiffVal and TDV, addressing the complexity of the tabulation method in what I hope is a useful and mathematically elegant manner.

I acknowledge Professor David W. Roberts for his thorough review of an earlier version of this article, his clarifications on the index notation, and his cordial exchange of long letters, which guided me toward a clearer, better-grounded manuscript. I also thank three anonymous reviewers, whose critiques, comments, and generous suggestions have enhanced the manuscript considerably.

T.M.H. was partially funded by the European Social Fund (POCH and NORTE 2020) and by National Funds (MCTES), through a FCT – Fundação para a Ciência e a Tecnologia postdoctoral fellowship (SFRH/BPD/115057/2016), as well as by National Funds, through the same foundation, under the project UIDB/04033/2020.

References

Aho K, Roberts DW, Weaver T (2008) Using geometric and non-geometric internal evaluators to compare eight vegetation classification methods. Journal of Vegetation Science 19: 549–562. https://doi.org/10.3170/2008-8-18406

Arbelaitz O, Gurrutxaga I, Muguerza J, Pérez JM, Perona I (2013) An extensive comparative study of cluster validity indices. Pattern Recognition 46: 243–256. https://doi.org/10.1016/j.patcog.2012.07.021

Armbruster WS (1995) The origins and detection of plant community structure: Reproductive versus vegetative processes. Folia Geobotanica et Phytotaxonomica 30: 483–497. https://doi.org/10.1007/BF02803978

Austin MP (2013) Vegetation and environment: discontinuities and continuities. In: van der Maarel E, Franklin J (Eds) Vegetation Ecology. John Wiley & Sons, Chichester, West Sussex, UK; Hoboken, NJ, US, 71–106. https://doi.org/10.1002/9781118452592.ch3

Barkman JJ (1989) Fidelity and character-species, a critical evaluation. Vegetatio 85: 105–116. https://doi.org/10.1007/BF00042260

Basseur M, Goëffon A (2015) Climbing combinatorial fitness landscapes. Applied Soft Computing 30: 688–704. https://doi.org/10.1016/j.asoc.2015.01.047

Belbin L (1987) The use of non-hierarchical allocation methods for clustering large sets of data. Australian Computer Journal 19: 32–41.

Bell ET (1934) Exponential polynomials. Annals of Mathematics 35(2): 258. https://doi.org/10.2307/1968431

Bishop CM (2006) Pattern recognition and machine learning. Springer, New York, US, 738 pp.

Bombina P, Tally D, Abrams ZB, Coombes KR (2024) SillyPutty: Improved clustering by optimizing the silhouette width. PLoS ONE 19: e0300358. https://doi.org/10.1371/journal.pone.0300358

Botta-Dukát Z (2008) Validation of hierarchical classifications by splitting dataset. Acta Botanica Hungarica 50: 73–80. https://doi.org/10.1556/ABot.50.2008.1-2.4

Bouveyron C, Celeux G, Murphy TB, Raftery AE (2019) Model-based clustering and classification for data science: with applications in R. Cambridge University Press, Cambridge, UK, 427 pp. https://doi.org/10.1017/9781108644181

Boyadzhiev KN (2012) Close encounters with the Stirling numbers of the second kind. Mathematics Magazine 85: 252–266. https://doi.org/10.4169/math.mag.85.4.252

Braun J, Furrer E (1913) Remarques sur l'étude des groupements de plantes. Bulletin de la Société Languedocienne de Geographie 36: 20–41. https://doi.org/10.3406/bulmi.1913.3540

Braun-Blanquet J (1921) Prinzipien einer Systematik der Pflanzengesellschaften auf floristischer Grundlage. Jahrbuches der St. Gallischen Naturwissenschaftlichen Gesellschaft 57: 305–351.

Braun-Blanquet J (1925) Zur Wertung der Gesellschaftstreue in der Pflanzensoziologie. Vierteljahrsschrift der Naturforschenden Gesellschaft in Zürich 70: 122–149.

Braun-Blanquet J (1932) Plant Sociology - The study of plant communities. McGraw-Hill Book Company, Inc., New York, US, 439 pp.

Bray JR, Curtis JT (1957) An ordination of the upland forest communities of Southern Wisconsin. Ecological Monographs 27: 325–349. https://doi.org/10.2307/1942268

Bruelheide H (2000) A new measure of fidelity and its application to defining species groups. Journal of Vegetation Science 11: 167–178. https://doi.org/10.2307/3236796

Bruelheide H, Flintrop T (1994) Arranging phytosociological tables by species-relevé groups. Journal of Vegetation Science 5: 311–316. https://doi.org/10.2307/3235854

Brun M, Sima C, Hua J, Lowey J, Carroll B, Suh E, Dougherty ER (2007) Model-based evaluation of clustering validation measures. Pattern Recognition 40: 807–824. https://doi.org/10.1016/j.patcog.2006.06.026

Cerdeira JO, Martins MJ, Silva PC (2012) A combinatorial approach to assess the separability of clusters. Journal of Classification 29: 7–22. https://doi.org/10.1007/s00357-012-9098-z

Češka A, Roemer H (1971) A computer program for identifying species-relevé groups in vegetation studies. Vegetatio 23: 255–277. https://doi.org/10.1007/BF02350624

Chytrý M, Tichý L, Holt J, Botta-Dukát Z (2002) Determination of diagnostic species with statistical fidelity measures. Journal of Vegetation Science 13: 79–90. https://doi.org/10.1111/j.1654-1103.2002.tb02025.x

Clements FE (1916) Plant succession: An analysis of the development of vegetation. Carnegie Institution of Washington, US, 512 pp. http://www.biodiversitylibrary.org/ia/cu31924000531818#page/1/mode/1up

Collins SL, Glenn SM, Roberts DW (1993) The hierarchical continuum concept. Journal of Vegetation Science 4: 149–156. https://doi.org/10.2307/3236099

Curran J, Williams A, Kelleher J, Barber D (2024) multicool: Permutations of Multisets in Cool-Lex Order. R package version 1.0.1. https://CRAN.R-project.org/package=multicool [accessed 30 May 2024]

Dale MB (1988) Knowing when to stop: Cluster concept - concept cluster. Coenoses 3: 11–32.

Dale MB (1995) Evaluating classification strategies. Journal of Vegetation Science 6: 437–440. https://doi.org/10.2307/3236243

Dale MB, Quadraccia L (1973) Computer assisted tabular sorting of phytosociological data. Vegetatio 28: 57–73. https://doi.org/10.1007/BF02386214

Dale MB, Webb LJ (1975) Numerical methods for the establishment of associations. Vegetatio 30: 77–87. https://doi.org/10.1007/BF02389609

De Cáceres M, Legendre P (2009) Associations between species and groups of sites: Indices and statistical inference. Ecology 90: 3566–3574. https://doi.org/10.1890/08-1823.1

De Cáceres M, Legendre P, Moretti M (2010) Improving indicator species analysis by combining groups of sites. Oikos 119: 1674–1684. https://doi.org/10.1111/j.1600-0706.2010.18334.x

Drummond C (2011) Classification. In: Sammut C, Webb GI (Eds) Encyclopedia of Machine Learning. Springer US, Boston, MA, US, 171–171. https://doi.org/10.1007/978-0-387-30164-8_111

Dufrêne M, Legendre P (1997) Species assemblages and indicator species: The need for a flexible asymmetrical approach. Ecological Monographs 67: 345–366. https://doi.org/10.2307/2963459

Ellenberg H (1956) Aufgaben und Methoden der Vegetationskunde. In: Walter H (Ed.) Grundlagen der Vegetationsgliederung, part 1. Einführung in die Phytologie. Eugen Ulmer, Stuttgart, 136.

Feoli E, Ganis P (2019) The use of the evenness of eigenvalues of similarity matrices to test for predictivity of ecosystem classifications. Mathematics 7: 245. https://doi.org/10.3390/math7030245

Feoli E, Ganis P (2021) Similarity, classification and diversity "an Eternal Golden Braid" in quantitative vegetation studies. Flora Mediterranea 31: 23–41. https://doi.org/10.7320/FlMedit31SI.023

Feoli E, Orlóci L (1979) Analysis of concentration and detection of underlying factors in structured tables. Vegetatio 40: 49–54. https://doi.org/10.1007/BF00052015

Gauch HG (1980) Rapid initial clustering of large data sets. Vegetatio 42: 103–111. https://doi.org/10.1007/BF00048877

Gauch HG (1982) Multivariate analysis in community ecology. 1st ed. Cambridge University Press, Cambridge, UK. https://doi.org/10.1017/CBO9780511623332

Goodall D (1953) Objective methods for the classification of vegetation. II. Fidelity and indicator value. Australian Journal of Botany 1(3): 434. https://doi.org/10.1071/BT9530434

Han J, Kamber M, Pei J (2012) Data mining: concepts and techniques. 3rd ed. Elsevier/Morgan Kaufmann, Waltham, MA, US, 703 pp.

Hennig C (2015) What are the true clusters? Pattern Recognition Letters 64: 53–62. https://doi.org/10.1016/j.patrec.2015.04.009

Hill MO (1979) TWINSPAN - A FORTRAN program for arranging multivariate data in an ordered two-way table by classification of the individuals and attributes. Section of Ecology and Systematics, Ithaca, New York. US.

Holzner W, Werger MJA, Ellenbroek GA (1978) Automatic classification of phytosociological data on the basis of species groups. Vegetatio 38: 157–164. https://doi.org/10.1007/BF00123266

Illyés E, Chytrý M, Botta-Dukát Z, Jandt U, Škodová I, Janišová M, Willner W, Hájek O (2007) Semi-dry grasslands along a climatic gradient across Central Europe: Vegetation classification with validation. Journal of Vegetation Science 18: 835–846. https://doi.org/10.1111/j.1654-1103.2007.tb02600.x

Jaccard P (1901) Distribution de la flore alpine dans le Bassin des Dranses et dans quelques régions voisines. Bulletin de la Société Vaudoise des Sciences Naturelles 37: 241–272.

Kaufman L, Rousseeuw PJ (1990) Finding groups in data: an introduction to cluster analysis. Wiley, New York, US, 342 pp. https://doi.org/10.1002/9780470316801

Kent M, Coker P (1992) Vegetation description and analysis - A practical approach. Belhaven Press, London, UK, 363 pp.

Koch LF (1957) Index of biotal dispersity. Ecology 38: 145–148. https://doi.org/10.2307/1932140

Kuhn N (1983) VEGTAB, ein Computer-Programm als Hilfe zur tabellarischen Vegetationsgliederung. Tuxenia: 499–522.

Lengyel A, Roberts DW, Botta-Dukát Z (2021) Comparison of silhouette-based reallocation methods for vegetation classification. Journal of Vegetation Science 32: e12984. https://doi.org/10.1111/jvs.12984

Lötter M, Mucina L, Witkowski E (2013) The classification conundrum: Species fidelity as leading criterion in search of a rigorous method to classify a complex forest data set. Community Ecology 14: 121–132. https://doi.org/10.1556/ComEc.14.2013.1.13

MacQueen J (1967) Some methods for classification and analysis of multivariate observations. In: Le Cam LM, Neymann J (Eds) Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. University of California Press, Berkeley, US, 281–297.

Madeira SC, Oliveira AL (2004) Biclustering algorithms for biological data analysis: A survey. IEEE/ACM Transactions on Computational Biology and Bioinformatics 1: 24–45. https://doi.org/10.1109/TCBB.2004.2

Maechler M, Rousseeuw P, Struyf A, Hubert M, Hornik K (2023) cluster: Cluster Analysis Basics and Extensions. R package version 2.1.6. https://CRAN.R-project.org/package=cluster [accessed 30 May 2024]

Mikolajczak A, Maréchal D, Sanz T, Isenmann M, Thierion V, Luque S (2015) Modelling spatial distributions of alpine vegetation: A graph theory approach to delineate ecologically-consistent species assemblages. Ecological Informatics 30: 196–202. https://doi.org/10.1016/j.ecoinf.2015.09.005

Milligan GW (1996) Clustering validation: results and implications for applied analyses. In: Arabie P, Hubert LJ, De Soete G (Eds) Clustering and Classification. World Scientific Publishing, River Edge, NJ, US, 341–375. https://doi.org/10.1142/9789812832153_0010

Monteiro-Henriques T, Cerdeira JO (2023) diffval: Vegetation Patterns. R package version 1.1.0. https://CRAN.R-project.org/package=diffval [accessed 30 May 2024]

Moravec J (1993) Syntaxonomic and nomenclatural treatment of Scandinavian-type associations and sociations. Journal of Vegetation Science 4: 833–838. https://doi.org/10.2307/3235622

Mucina L (1997) Classification of vegetation: Past, present and future. Journal of Vegetation Science 8: 751–760. https://doi.org/10.2307/3237019

Mucina L, van der Maarel E (1989) Twenty years of numerical syntaxonomy. Vegetatio 81: 1–15. https://doi.org/10.1007/BF00045509

Mueller-Dombois D, Ellenberg H (1974) Aims and methods of vegetation ecology. John Wiley & Sons, New York, US.

Murphy AH (1996) The Finley affair: A signal event in the history of forecast verification. Weather and Forecasting 11: 3–20. https://doi.org/10.1175/1520-0434(1996)011%3C0003:TFAASE%3E2.0.CO;2

Oksanen J, Hill MO (2025) twinspan: Two-Way Indicator Species Analysis. R package version 0.9-5. https://github.com/jarioksa/twinspan [accessed 12 Jan 2025]

Oksanen J, Simpson GL, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'Hara RB, Solymos P, Stevens MHH, … Borman T (2024) vegan: Community Ecology Package. R package version 2.6-8. https://CRAN.R-project.org/package=vegan [accessed 12 Jan 2025]

Orlóci L (1975) Multivariate analysis in vegetation research. Springer Netherlands, Dordrecht, NL, 276 pp. https://doi.org/10.1007/978-94-017-5608-2

Pakgohar N, Lengyel A, Botta-Dukát Z (2024) Quantitative evaluation of internal cluster validation indices using binary data sets. Journal of Vegetation Science 35: e13310. https://doi.org/10.1111/jvs.13310

Peet RK, Roberts DW (2013) Classification of natural and semi-natural vegetation. In: van der Maarel E, Franklin J (Eds) Vegetation ecology. John Wiley & Sons, Chichester, West Sussex, UK; Hoboken, NJ, US, 28–70. https://doi.org/10.1002/9781118452592.ch2

Pillar VD (1999) How sharp are classifications? Ecology 80: 2508–2516. https://doi.org/10.1890/0012-9658(1999)080[2508:HSAC]2.0.CO;2

Podani J (1986) Comparison of partitions in vegetation studies. Abstracta Botanica 10: 235–290.

Podani J, Feoli E (1991) A general strategy for the simultaneous classification of variables and objects in ecological data tables. Journal of Vegetation Science 2: 435–444. https://doi.org/10.2307/3236025

Poore MED (1955) The use of phytosociological methods in ecological investigations: II. Practical issues involved in an attempt to apply the Braun-Blanquet system. Journal of Ecology 43: 245–269. https://doi.org/10.2307/2257133

Poore MED (1956) The use of phytosociological methods in ecological investigations: IV. General discussion of phytosociological problems. Journal of Ecology 44: 28–50. https://doi.org/10.2307/2257153

Popma J, Mucina L, Tongeren O, van der Maarel E (1983) On the determination of optimal levels in phytosociological classification. Vegetatio 52: 65–75. https://doi.org/10.1007/BF00037763

R Core Team (2024) R: A language and environment for statistical computing. R version 4.4.2. https://www.R-project.org/ [accessed 3 Nov 2024]

Roberts DW (2015) Vegetation classification by two new iterative reallocation optimization algorithms. Plant Ecology 216: 741–758. https://doi.org/10.1007/s11258-014-0403-2

Roberts DW (2020) optpart: Optimal partitioning of similarity relations. R package version 3.0-3. https://CRAN.R-project.org/package=optpart [accessed 31 May 2024]

Roberts DW (2023) labdsv: Ordination and multivariate analysis for ecology. R package version 2.1-0. https://CRAN.R-project.org/package=labdsv [accessed 31 May 2024]

Roleček J, Tichý L, Zelený D, Chytrý M (2009) Modified TWINSPAN classification in which the hierarchy respects cluster heterogeneity. Journal of Vegetation Science 20: 596–602. https://doi.org/10.1111/j.1654-1103.2009.01062.x

Rousseeuw PJ (1987) Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 20: 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Schmidtlein S, Tichý L, Feilhauer H, Faude U (2010) A brute-force approach to vegetation classification. Journal of Vegetation Science 21: 1162–1171. https://doi.org/10.1111/j.1654-1103.2010.01221.x

Schmidtlein S, Collison J, Pfannendoerfer R, Tichý L (2025) isopam: Clustering of sites with species data. R package version 3.2. https://CRAN.R-project.org/package=isopam [accessed 16 Jan 2025]

Sokal RR, Sneath PH (1963) Principles of numerical taxonomy. W.H. Freeman and Company, San Francisco, US, and London, UK, 359 pp.

Spatz G (1972) Eine Möglichkeit zum Einsatz der elektronischen Datenverarbeitung bei der pflanzensoziologischen Tabellenarbeit. In: van der Maarel E, Tüxen R (Eds) Grundfragen und Methoden in der Pflanzensoziologie/Basic Problems and Methods in Phytosociology. Bericht Internationale Symposium Rinteln 1970, 251–258. https://doi.org/10.1007/978-94-015-7241-5_18

Thiergart T, Schmitz U, Landan G, Martin WF, Dagan T (2014) Application and comparative performance of network modularity algorithms to ecological communities classification. Acta Societatis Botanicorum Poloniae 83: 93–102. https://doi.org/10.5586/asbp.2014.016

Tichý L, Chytrý M (2006) Statistical determination of diagnostic species for site groups of unequal size. Journal of Vegetation Science 17: 809–818. https://doi.org/10.1111/j.1654-1103.2006.tb02504.x

Tichý L, Chytrý M, Hájek M, Talbot SS, Botta-Dukát Z (2010) OptimClass: Using species-to-cluster fidelity to determine the optimal partition in classification of ecological communities. Journal of Vegetation Science 21: 287–299. https://doi.org/10.1111/j.1654-1103.2009.01143.x

Tüxen R, Ellenberg H (1936) Der systematische und der ökologische Gruppenwert. Jahresbericht der Naturhistorischen Gesellschaft zu Hannover 81–87: 171–184.

van der Maarel E (1979) Transformation of cover-abundance values in phytosociology and its effects on community similarity. Vegetatio 39: 97–114. https://doi.org/10.1007/BF00052021

van der Maarel E, Janssen JGM, Louppen JMW (1978) TABORD, a program for structuring phytosociological tables. Vegetatio 38: 143–156. https://doi.org/10.1007/BF00123265

Ward JH (1963) Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association 58: 236–244. https://doi.org/10.1080/01621459.1963.10500845

Westhoff V, van der Maarel E (1978) The Braun-Blanquet approach. In: Whittaker RH (Ed) Classification of plant communities. Dr. W. Junk, The Hague, NL; Boston, US, 287–399. https://doi.org/10.1007/978-94-009-9183-5_9

Whittaker RH (1962) Classification of natural communities. Botanical Review 28: 1–239. https://doi.org/10.1007/BF02860872

Wildi O (1989) A new numerical solution to traditional phytosociological tabular classification. Vegetatio 81: 95–106. https://doi.org/10.1007/BF00045515

Willner W (2006) The association concept revisited. Phytocoenologia 36: 67–76. https://doi.org/10.1127/0340-269X/2006/0036-0067

Willner W, Tichý L, Chytrý M (2009) Effects of different fidelity measures and contexts on the determination of diagnostic species. Journal of Vegetation Science 20: 130–137. https://doi.org/10.1111/j.1654-1103.2009.05390.x

Wilson JB (2012) Species presence/absence sometimes represents a plant community as well as species abundances do, or better. Journal of Vegetation Science 23: 1013–1023. https://doi.org/10.1111/j.1654-1103.2012.01430.x

Witten IH (2009) Classification. In: Liu L, Özsu MT (Eds) Encyclopedia of Database Systems. Springer US, Boston, MA, US, 331–335. https://doi.org/10.1007/978-0-387-39940-9_552

Appendix 1. Terminological and epistemological considerations

Initial remarks

Most of the issues discussed in this appendix stem from correspondence with Professor David W. Roberts, who kindly exchanged ideas with me after the submission of the first version of this article. Anticipating that these issues would likely be raised by other researchers, Professor Roberts recommended their discussion. I summarize some of the viewpoints that I have defended, highlighting subtleties that I believe merit consideration and that underpin the development of this method.

DiffVal, TDV, and TDV-optimization have the potential to reignite long-standing debates on the nature of vegetation communities. The use of the word 'differential' in their names may also resurrect old terminological discussions. Precisely because such debates have persisted for over a century, without reaching a clear consensus among vegetation scientists, I intentionally avoided delving into these epistemological or terminological matters in the body of the article. The sole objective of the article is to present a method for vegetation classification, using a block-based approach. As with any other method, the suitability of the approach for the intended analysis must be assessed by the user. Furthermore, the article makes no claim to resolve long-standing issues about the nature of vegetation communities.

Going back to the seminal concepts of vegetation science

Characteristic species, fidelity, and differential species: The notion of exclusiveness

At the beginning of the 20th century, characteristic species were those that indicated a particular environment or site. To refine this concept, Braun and Furrer (1913, p. 21) proposed that characteristic species should be those located exclusively, or almost exclusively, in a given plant community. Shortly thereafter, the notion of species fidelity to a plant community emerged, and around 1918, Braun-Blanquet proposed five degrees of fidelity (see Braun-Blanquet 1921, p. 316). Braun-Blanquet emphasized the role of faithful species in distinguishing and demarcating plant communities, using the first three degrees of fidelity to define three types of characteristic species: exclusive (*exclusif*), elective (*électif*), and preferring (*préférant*) (see

Braun-Blanquet 1921, p. 316). These were later renamed: exclusive (*treue*), selective (*feste*), and preferential (*holde*) (Braun-Blanquet 1932, p. 59).

The concept of fidelity relies on how the species are distributed across different plant communities. Braun-Blanquet refers to this as the "sociological distribution of species" (Braun-Blanquet 1932, p. 58), which is determined by examining (i) the distribution of the species presences among plant communities, or (ii) the distribution of species abundance or vitality across those communities.

The first three degrees of fidelity imply that (i) all, or almost all, of the presences are confined to a ***single*** plant community; (ii) a higher frequency is confined to a ***single*** plant community; or (iii) the higher abundances or vitalities are confined to a ***single*** plant community. In short, when any of these three conditions hold, the species is said to be centred in that community (see Whittaker 1962, p. 132).

Since the early stages of proposition of the phytosociological method and before the notions of fidelity and differential species emerged, Braun-Blanquet and Furrer were well aware that characteristic species were not enough to differentiate all plant communities (Braun and Furrer 1913, p. 26). In fact, a few years later, in 1925, the idea that some other species could be used for the sharp differentiation of plant communities, without requiring such strict confinement within a single community, matured into the concept of differential species. According to Braun-Blanquet (1925, p. 134), the term was proposed together with Walo Koch.

Differential species are those appearing "only in one of two or more related societies" (Braun-Blanquet 1925, p. 134, 1932, p. 59), i.e., a concept clearly grounded in the notion of exclusiveness. Not by chance the most faithful species are called exclusive species by Braun-Blanquet. Also not by chance, the word exclusiveness is given as a synonym of fidelity by Whittaker (1962, p. 132) and by Mueller-Dombois and Ellenberg (1974, p. 203). All these concepts (characteristic species, fidelity, and differential species) are grounded in the exclusiveness notion, irrespective of their different purposes and utilities.

The loss of the characteristic species at the association level

Braun-Blanquet presents fidelity as a pivotal concept in the study of plant communities at the association level. For him, fidelity is a synthetic quality of plant communities, determined after the associations are established by tabulating each association side by side (Braun-Blanquet 1932, pp. 52, 58).

The initial community-unit hypothesis put forward by Clements (1916), contemporary with Braun-Blanquet's early texts on characteristic species, proposed a strong overlap of species within a given community and almost perfect non-overlap with species from neighbouring communities (see, e.g., Collins et al. 1993 for a graphical representation). At that time, it was believed that at least one exclusive species could be found for most associations, in a univocal manner. If such a clear relationship between an association and a single exclusive (*exclusif/treue*) species truly existed, delimiting associations would also entail identifying their exclusive species, i.e., associations would be delimited simultaneously with the identification of first-degree fidelity. Naturally, this would require comprehensive knowledge of flora distribution. The second and third degrees of fidelity could be determined subsequently.

From the 1940s to the 1960s, the narrowing of the association concept, as described in Mueller-Dombois and Ellenberg (1974, p. 206), implied that such a univocal relationship could not exist in practice at the association level. Eventually, after considerable discussions during the 20th century (see e.g. Whittaker 1962; Westhoff and van der Maarel 1978), the idea of globally (or even regionally) valid characteristic species at the association level was considered unfeasible. Likewise, the notion that the higher frequencies (or higher abundances or vitalities) could be confined to a ***single*** plant community was also considered unfeasible, as a species may perform well in more than one association. Species are simply not centred within the contemporary conception of plant communities.

At this stage, it no longer made sense to refer to characteristic species at the association level. In fact, Poore (1956) and Moravec (1993) highlighted that, in 1941, Gams asserted there is no fundamental difference between characteristic and differential species, thus triggering "the decline of the character species doctrine" (Moravec 1993). Furthermore, Moravec notes that characteristic species were frequently reassessed as differential species, which led to the introduction of the neutral term 'diagnostic species' by Whittaker in 1962.

Differential species, as initially defined, became crucial in the delimitation of associations. Eventually, the floristic distinction between associations was based entirely on differential species. In practice, phytosociologists find combinations of species, with these species potentially serving as differential species in other associations (Ellenberg 1956). Without the univocal relationship, the complexity increases considerably, and first-degree fidelity can no longer be determined simultaneously with the delimitation of the association.

Even though the narrowing of the association concept compromised the identification of characteristic species at the association level, there was considerable reluctance to abandon the doctrine in the following decades (Moravec 1993). After the 1960s, several authors proposed adjustments to the original concepts, possibly in an attempt to salvage the notion of characteristic species at the association level. Moreover, as differential species came to be seen as essential for delimiting associations – yet clearly dependent on the context of comparison – some authors proposed refinements to the concept. Even if subtle, changes in the concepts of fidelity and differential species over the years have had significant consequences when applied in numerical approaches. Therefore, I believe a full exposition is necessary for clarity.

Changes in postulates over the years

Nuances still under the exclusiveness notion

Ellenberg (1956) illustrates the practice of phytosociology, particularly the tabulation technique, detailing how it relies on identifying co-occurrence patterns of differential species. This practical exemplification is available in English in Mueller-Dombois and Ellenberg (1974).

Ellenberg introduces a slightly modified concept of differential species. While Braun-Blanquet and Walo Koch define differential species as species appearing "only in one of two or more related societies", Ellenberg's approach allows a differential species to occur in more than one vegetation unit, provided it is absent from at least one of the units under comparison. In other words, the species remains exclusive to one or some of the emerging groups. This adjustment is particularly relevant to the present work, as it is incorporated into the DiffVal formulation.

A shift from the exclusiveness notion to the concentration notion

Whittaker (1962, p. 132), while discussing the application of the phytosociological method, makes no significant changes to Braun-Blanquet's concept of fidelity. However, on pp. 130–131, Whittaker associates the identification of characteristic species with the specific set of relevés under tabulation – a much more limited context compared to Braun-Blanquet's broader approach. When addressing differential species (p. 131), Whittaker introduces some modifications:

> "*If there are no character-species for a grouping; but certain species are present in samples of this grouping and absent* ***or clearly less important*** *in those of another, closely related grouping, these are* differential-species *for a unit* ***usually of lower rank than the association.***" [my own bold emphasis]

Compared to Braun-Blanquet's initial definition, there are two modifications: (i) Whittaker's retains the possibility that a species may be entirely absent from one grouping but relaxes this criterion by adding: "or clearly less important". In contrast, Braun-Blanquet strictly required complete absence from one (or more) of the compared groups. (ii) Differential species are now considered useful for distinguishing units below the association level, whereas Braun-Blanquet also applied the concept at the association level (note that, at Braun-Blanquet's time, the association concept was even broader).

Westhoff and van der Maarel (1978, p. 324) later defined fidelity as "the degree to which a species is ***concentrated in one***" syntaxon. They further described differential species as those that "distinguish two closely related syntaxa by presence in most samples of the one and ***absence in most samples*** of the other" (p. 372). This definition fully relaxes the requirement that differential species be completely absent from one or more compared plant communities. Moreover, they restrict the use of differential species to ranks below the association level.

The concentration notion was also employed by Feoli and Orlóci (1979) and Podani and Feoli (1991), in their analysis of concentration. Neither of these works refers to fidelity or differential species, although the latter briefly mentions tabular sorting.

Barkman (1989, p. 105) defines fidelity as "the degree of preference of a species for a given association". By introducing the concept of preference, Barkman's definition moves away from restricting fidelity to a single plant community, giving it a continuous, statistical flavour. In p. 106, Barkman refers to a "wide sense" definition of differential species, which is commonly used in practice (the one I follow in the article body). He also provides a "more strict and formal" definition: "taxa that have an optimum i[n] only ***one*** syntaxon within the next higher syntaxon". The introduction of the concept of optimum aligns with the concentration notion.

Barkman extends the concept of fidelity to that of differential species, asserting that this extension is implicit in a work by Koch in 1926, and stating that "there is no reason not to use the fidelity classes for differential taxa as well". At that time, and still today, differential species form the foundation of vegetation classification, as evidenced by the works of Ellenberg (1956) and Mueller-Dombois and Ellenberg (1974). Therefore, extending the fidelity classes to differential taxa/species would be epistemologically catastrophic. Barkman's formal definition of differential species, if accepted, would cripple their use in the delimitation of plant communities. For instance, at the association level, Barkman's formal definition requires knowing the alliance circumscription before establishing the associations. However, an alliance can only be established after the associations have been determined and compared, generating a clear paradox.

Dufrêne and Legendre (1997) introduced an ecological perspective to fidelity, describing it as "based upon ecological preferences". Similarly, they defined differential species as "species with clear ecological preferences". These definitions were inspired by Hill's (1979) formulation of the TWINSPAN procedure. While Dufrêne and Legendre's definitions are not far from Braun-Blanquet's original concepts, their proposal of quantification differs significantly. They suggest that the relative frequency of presences in the vegetation cluster (quantity B) serves as a measure of fidelity. However, on closer examination, only quantity A conceptually aligns with a fidelity measure, as it evaluates the accumulation of abundance in one group ***relative to other groups*** (with abundance often replaced by cover in vegetation science). Quantity B, on the other hand, is not relative to other groups, and, therefore, cannot accurately represent a "preference". This distinction has been highlighted previously (see Tichý and Chytrý 2006; De Cáceres and Legendre 2009; Willner et al. 2009). Nevertheless, quantity A (or the IndVal, A × B, as a whole) still presents a notable difference when compared to Braun-Blanquet's concept of differential species, a difference I will revisit further below.

Bruelheide (2000) provides a definition of fidelity, again relying on the concentration notion:

> *"Fidelity is one of the most important concepts of the Braun-Blanquet (Zürich-Montpellier) approach. Generally speaking, fidelity is the degree to which a species is* ***concentrated*** *in a given vegetation unit. The fidelity of a species determines whether it can be considered a differential or character species or just a companion or accidental species. A character species can be interpreted as a special case of a differential species: a differential species shows a distinct* ***accumulation of occurrences*** *in one or more vegetation units; whereas a character species should accumulate in only one single vegetation unit (Barkman 1989: 106)."* [my own bold emphasis]

Bruelheide's definition of fidelity likely draws support from the previously mentioned works, such as the review by Westhoff and van der Maarel (1978), though references were not provided since fidelity was described in general terms. It is worth noting that the accumulation of abundance, as measured by Dufrêne and Legendre's (1997) quantity A, is analogous to the "accumulation of occurrences" mentioned by Bruelheide. Toward the end of the paragraph, Bruelheide links this fidelity concept with a modified version of the differential species concept, based on Barkman's (1989) definition. Bruelheide adapted Barkman's notion of "optimum" into "accumulation of occurrences".

Considering Bruelheide's work, I highlight three issues:

(i) The fundamental shift from Braun-Blanquet's original concept: In 1925, differential species were those ***accumulating absences*** in one or more vegetation units (i.e., being completely absent from all but one vegetation unit). In contrast, by 2000, differential species were those ***accumulating presences*** in one or more vegetation units (and possibly present in all vegetation units).

(ii) Bruelheide raises concerns about circular reasoning in his own work. However, the primary cause of this circularity stems from a change he made to the definition of fidelity. Since Bruelheide uses the concept to assist in classification, fidelity is defined in relation to any "vegetation unit" – not necessarily ranked units (syntaxa). This represents a significant departure from the long-established phytosociological tradition, where fidelity was studied and determined after the classification process (see Table A1). This decision clearly introduces circularity into the fidelity concept.

(iii) Bruelheide's concerns about circularity are not satisfactorily addressed. The author claims that circularity is avoided by optimizing the *u* measure. However, optimization does not offer an epistemological solution to this circularity, as the definition of fidelity remains circular in relation to the definition of vegetation unit (Bruelheide 2000, p. 167), regardless of the optimization process.

These issues, particularly (i) and (ii), are deeply rooted in the fact that Bruelheide built on Barkman's idea of extending the fidelity concept to the differential species concept.

As we can see, the exclusiveness notion found in the works of Braun-Blanquet, Whittaker, and Mueller-Dombois and Ellenberg progressively mutated into the idea of concentration, which seems to have become well-established in vegetation science. Chytrý et al. (2002) described fidelity (also in general terms) as "a measure of species concentration in vegetation units" and, building upon Bruelheide's (2000) *u* proposal, they illustrate the use of the phi coefficient. As a side note, it should be mentioned that the square value of the phi coefficient was first proposed in the meteorology field by Doolitte in 1885 (see Murphy 1996).

Later, De Cáceres and Legendre (2009) provided the same definition as Chytrý et al. (2002) (also in general terms and without references), while exploring variants of the IndVal and the phi coefficient.

Tables A1, A2 summarize the changes in the concepts over the years, highlighting the subtle differences between the two paradigms: exclusiveness (in light grey) and concentration (in dark grey).

Exclusiveness versus concentration: What are the consequences?

In the previous subchapters, I have examined the emergence of the definitions of fidelity and differential species, as well as their changes over the years, highlighting the existence of two paradigms: one based on exclusiveness and the other on concentration.

Concentration-based fidelity is now used in the examination of syntaxa after a classification has been obtained, rather than during the classification process itself. This is indisputable, as previous attempts to use fidelity directly in classification have led researchers into paradoxes, circular definitions, and begging-the-question misconceptions. Fidelity is derived after the different syntaxa are established (Chytrý et al. 2002), a principle that holds true under both paradigms.

I will continue exploring the fidelity concept here, as some readers might find it interesting to highlight the differences between DiffVal and other indices currently used in the study of concentration-based fidelity. However, I caution the reader that DiffVal, TDV and TDV-optimization are grounded in the exclusiveness-based differential species concept, as implemented in Ellenberg (1956) and in Mueller-Dombois and Ellenberg (1974). They are not based on any of the concentration-based fidelity concepts, even though they are related. TDV-optimization is focused on the floristic delimitation of plant communities (i.e., vegetation classification, particularly at the association level), rather than the study of species fidelity after a classification is completed.

If we closely examine the first three degrees of the original fidelity concept (Braun-Blanquet 1932, p. 59; Mueller-Dombois and Ellenberg 1974, p. 203), we observe a progressive relaxation of the exclusiveness. The

Table A1. The concepts of fidelity (and characteristic species) and their nuances over the years (relevant subtleties in italics). When authors provide numerical approaches to these concepts, the characteristics of such approaches are also considered in the description of the concepts' intensions presented in the table.

			Fidelity and character species intensions[1]							
			Species are required to occur exclusively in *one to some* communities.		Species may occur in *several* communities, but the higher abundance/cover or vitalities are found in *one*...	The exclusiveness requirement is fully relaxed. Species presences/cover are concentrated or accumulated in *one or more than one*...				
Year	**Author(s)**	**Number of levels of characteristic species**	**Species presences exclusive or almost exclusive of *one*...**	**A higher frequency of the species is found in *one*...**			**Geographical context**	**Floristic-vegetational context**	**Obtained...**	**Used...**
1913	Braun-Blanquet and Furrer	1	...association.	---	---	---	≈ Regional (not clearly stated).	As defined by the regional context. (not clearly stated).	...simultaneously to classification.[2]	...to demarcate the association[3] level.
1918 / 1932	Braun-Blanquet	3	...syntaxon.	...syntaxon.	...syntaxon.	---	Regional.	As defined by the regional context.	...after classification.	...during the ranking of the vegetation units (syntaxonomic positioning).
1956	Ellenberg	3	...syntaxon.	...syntaxon.	...syntaxon.	---	Local, regional, and absolute.	As defined by the regional context.	...after classification.	...in ranking (syntaxonomic positioning).
1962	Whittaker	Not relevant	...syntaxon[4]	...syntaxon[4]	...syntaxon[4]	---	As defined by the set of samples/relevés being analysed.	Set of samples/relevés.	Not clear.[5]	...to determine the complete characteristic species-combination, i.e. classification; but also used to characterize the different syntaxonomic ranks.[5]
1974	Mueller-Dombois and Ellenberg	3	...syntaxon.	...syntaxon.	...syntaxon.	---	Local, regional, and absolute.	As defined by the regional context.	...after classification.	...during the ranking of the vegetation units (syntaxonomic positioning).
1978	Westhoff and van der Maarel	3	...syntaxon.	...syntaxon.	...syntaxon.	---[6]	Local, regional, and general.	As defined by the regional context.	...after classification.	...at the end of the synthetical phase, preparing the syntaxonomic phase.
1989	Barkman	---[7]	---[7]	---[7]	---[7]	...syntaxon.[8]	Regional.	As defined by the regional context.[9]	...after classification.	...to express the ideas of optima and role of species in plant communities (or other elementary coenological units).
1997	Dufrêne and Legendre	Not relevant	---	---	---	... groups of a certain classification.	As defined by the set of samples/relevés being analysed.	Not relevant	...after classification.	...to explore the bond between species and the groups of a certain classification.
2000	Bruelheide	Not relevant	---	---	---	...vegetation unit	As defined by the set of samples/relevés being analysed.	Set of samples/relevés.	...during classification.[10]	...to demarcate vegetation units.[10]

[1] In logic, intension refers to the internal content of a concept. [2] When a univocal relationship was expected between *exclusif/treue* species and associations, obtaining such characteristic species was simultaneous to the classification process. [3] Earlier, the concept of association seemed much broader than later in the 20th century. Therefore, "association" could easily correspond to, for example, modern alliances (see Mueller-Dombois and Ellenberg 1974). [4] While describing the formal classification in the Braun-Blanquet school, Whittaker (p. 130–131) defines characteristic species as those "largely confined to a grouping" or "clearly less important in the rest of the sample material". On p. 132, he describes "fidelity or exclusiveness (*Treue*)" as "the degree to which a species is restricted to a particular community-type, or is centered in that type and of less importance or vitality in other types". These definitions closely align with those originally proposed by Braun-Blanquet, so the exclusiveness requirement appears to be accepted by Whittaker. However, Whittaker exemplifies the determination of characteristic species during basic relevé classification (p. 130–131), which diverges from the practice under Braun-Blanquet's school. In practice, Whittaker treats characteristic species as regular differential species confined to a single grouping (among those being tabulated), while he defines differential species as those that might occur in more than one grouping (among those being tabulated) but are still useful for pairwise comparisons of similar groupings. [5] Whittaker is not entirely clear on this. On pp. 130–131, characteristic species seem to be obtained during the classification process, but on pp. 131–132, he refers to the study of fidelity as a synthetic character determined after "the body of samples has been grouped into community-types". [6] While Westhoff and van der Maarel (1978) use the notion of concentration when describing fidelity, they still adhere to Braun-Blanquet's original concept of characteristic species, ensuring that such concentration occurs only in one syntaxon. Note that in the same work, the authors relax the exclusiveness requirement for differential species (see Table A2). [7] Barkman (1989) refers to the three levels of characteristic species, but focuses on operationalizing the fidelity concept numerically, based on the notion of concentration and building on Szafer & Pawłowski previous work. [8] Barkman (1989) defines characteristic species as those with an optimum in one syntaxon. However, his proposed operationalization of fidelity is a pairwise index (TC'R), which, in practice, would allow species to show an optimum in more than one syntaxon. [9] While applying the fidelity concept to differential species, Barkman proposes using the next higher syntaxon as the floristic-vegetational context (see Table A2). [10] Bruelheide relies on Barkman's definitions of fidelity, expanding it to any "vegetation unit", i.e., not necessarily ranked units (syntaxa). This departs from the long-standing tradition of obtaining characteristic species only after classification. This change introduces a circular argument, which Bruelheide identifies. Barkman (1989) also proposed expanding the use of fidelity to differential species (see Table A2), but without operationalizing or illustrating its use in classification. Bruelheide operationalizes it, proposing the application of the fidelity concept to vegetation classification. However, the use of the *u* as a fidelity measure relaxes the exclusiveness requirement of traditional tabulation (see Table A2, and the Appendix 1 text) without solving the identified circularity.

Table A2. The concept of differential species and its nuances over the years (relevant subtleties in italics). When authors provide numerical approaches to this concept, the characteristics of such approaches are also considered in the description of the concept's intensions presented in the table.

Year	Author(s)	Differential species intensions[1]			Obtained...	Used...
		Floristic-vegetational context	Considering the given floristic/vegetational context...			
			...differential species presences are *exclusive to*...	...differential species presences/cover are *concentrated but not necessarily exclusive to*...		
1925	Braun-Blanquet and Walo Koch	*Two or more* similar vegetation units.	...*one* of the vegetation units under comparison. (N.b., cannot occur in all of them).	---	...during classification.	...to define associations and lower ranks, i.e., used in classification.[2]
1956	Ellenberg	The set of relevés under analysis. (Usually, a set of similar relevés, potentially containing two or more vegetation units).	...*one or more than one* of the emerging vegetation units. (N.b., cannot occur in all of them).	---	...during classification (more concretely during tabulation).	...to define unranked vegetation units. (I.e., used in classification, specifically through tabulation.)
1962	Whittaker	*Two* closely related vegetation units, among the sampled plots (pairwise comparisons among the emerging groups).	---	...*one* vegetation unit. (N.b., possibly also occurring in the second unit under comparison).	...during classification.	...to define lower ranks than association, i.e., used in classification.
1974	Mueller-Dombois and Ellenberg	The set of relevés under analysis. (Usually, a set of similar relevés, potentially containing two or more vegetation units).	...*one or more than one* of the emerging vegetation units. (N.b., cannot occur in all of them.)	---	...during classification (more concretely during tabulation).	...to define unranked vegetation units. (I.e., used in classification, specifically through tabulation.)
1978	Westhoff and van der Maarel (These authors distinguish *differential* and *differentiating* species)	For *differentiating* species: The set of relevés under analysis. (As defined by Ellenberg, see above in this table.)	...*one or more than one* of the emerging vegetation units. (As defined by Ellenberg)	---	...during classification. (As defined by Ellenberg)	...to define unranked vegetation units. (As defined by Ellenberg)
		For *differential* species: *Two* closely related syntaxa.	---	...*one* vegetation unit. (N.b., possibly also occurring in the second unit under comparison).	...while comparing two syntaxa.	...to define lower ranks than association, i.e. used in classification.
1989	Barkman	The next higher syntaxon, thus *two or more* vegetation units.	---	...*one or more than one* vegetation unit(s) (N.b., possibly also occurring in all the other units under comparison).	Not clear.[3]	Not clear.[3]
1997	Dufrêne and Legendre	Sites/relevés under analysis.	---	...*one or more than one* vegetation unit(s) (N.b., possibly also occurring in all the other units under comparison).	...after classification.[4]	...to explore the bond between species and the groups of a certain classification. [4]
2000	Bruelheide	The next higher syntaxon.	---	...*one or more than one* vegetation unit(s) (N.b., possibly also occurring in all the other units under comparison).	...by optimization of indices.[5]	...to assist the extraction of differential species groups, i.e. to assist classification.[6]
2025	(the current work)	The set of relevés under analysis. (Usually, a set of similar relevés, potentially containing two or more vegetation units).	...*one or more than one* of the emerging vegetation units. (N.b., cannot occur in all of them).	---	...by optimization of TDV.	...to define unranked vegetation units. (i.e., used in classification, specifically through tabulation).

[1] In logic, intension refers to the internal content of a concept. [2] Earlier, the concept of association seemed much broader than later in the 20th century. Therefore, in this context, "associations and lower ranks" could easily correspond to, e.g., alliances, associations and subassociations (see Mueller-Dombois and Ellenberg 1974). [3] Barkman proposes extending fidelity measures to differential species. This would hinder the use of the differential species, as defined by Barkman, in classification, as it introduces a circular definition. See Appendix 1 text. [4] Dufrêne and Legendre (1997) present IndVal as independent of any classification method, including clustering procedures based solely on floristic data. This view is revised in De Cáceres and Legendre (2009) and De Cáceres et al. (2010), where the authors acknowledge that, if the classification is based on species composition, the independence no longer holds. Nowadays, IndVal is considered suitable for exploring the relationship between species and ecologically defined groups. The initial purposes of the method (typology comparison, serving as a stopping rule in cluster splitting, or assisting hierarchical classification of sites) have seldom been implemented by vegetation scientists. [5] Note that the use of optimization does not guarantee the independence between the obtained vegetation units and their respective fidelity indices. [6] Note that Bruelheide relies on Barkman's floristic-vegetational context (i.e., the next higher syntaxon), which introduces a circular definition. Optimization does not offer an epistemological solution for this circularity, as the definition remains circular regardless. Additionally, the use of the *u* measure as a fidelity index relaxes the exclusiveness requirement of traditional tabulation (see the Appendix 1 text).

first degree is strongly rooted in the exclusiveness of species presences to ***a single vegetation unit***, relaxing this notion only slightly by stating that a species is "completely or almost completely confined to one community" or "exclusively or almost exclusively found only in a single association". The second degree further relaxes the exclusiveness of species presences, allowing for presence in more communities, but not all. Note that "other" associations/communities are mentioned but not "several", as in the third degree. Still, a higher frequency is expected in ***one vegetation unit*** only. The third degree relaxes exclusiveness the most: the species is present in "several" associations/communities, and the focus shifts to the species' "better vitality" or being "optimally developed" or "abundant" in ***one vegetation unit***.

The first two degrees of the original fidelity concept can also be described as a concentration or accumulation of presences in one (or a few) groups, ***with exclusiveness to that group (or those groups) guaranteed.*** This is where we begin to spot some disparities: IndVal, the phi coefficient, and similar indices ***do not ensure exclusiveness to the groups*** and can return intermediate to high values for species with presences in all the groups being compared.

Table A3 shows how the commonly used group-equalized versions of IndVal ($\mathrm{IndVal}^{g}_{ind}$), the phi coefficient ($r^{g}_{\phi}$), and the point-biserial correlation coefficient (r^{g}_{pb}) behave based on changes in (i) exclusiveness to groups, (ii) concentration of presences, and (iii) concentration of cover. Calculations for individual groups and for the best group combination are presented, obtained using the "multipatt" function from the R package *indicspecies* (De Cáceres and Legendre 2009). DiffVal is also included for comparison. From the table, we observe that IndVal, the phi coefficient, and the point-biserial correlation coefficient – while used for the study of concentration-based fidelity – are grounded in a notion that differs essentially from the original definition provided by Braun-Blanquet. The exclusiveness rationale behind the original definition is particularly straightforward. For example, considering Table A3, what would be the diagnostic value of species 5 in the field? Would its presence or absence in a stand allow us to differentiate any of the three groups? Similarly for species 8, 9, and 10, would their presence in the field, say with a cover-abundance value of 2 or 3, allow us to differentiate any of the three groups? The answer is certainly no!

Species 5 and 8 demonstrate how the phi-coefficient returns reasonably high values, detecting the concentration of the presences in one of the groups, even though exclusiveness to that group is not ensured. When considering the best group combination, species 4, 7 and 9 show cases where the phi-coefficient behave quite differently from DiffVal. Similarly, the high values of IndVal and the point-biserial correlation coefficient for species 8, 9 and 10 illustrate how these indices detect the concentration of cover in one group, again irrespective of exclusiveness. Note that DiffVal is zero for all three species. When considering the best group combination, species 4, 7, 9 and 10 provide examples where IndVal and the point-biserial correlation coefficient can diverge significantly from DiffVal.

The third degree of the original fidelity concept is the closest to a general notion of concentration, as it is postulated almost independently of exclusiveness (i.e., it assumes the species occurs in "several" associations/communities – though not necessarily in all of them). The third degree is also the closest to the quantity A used in IndVal. However, it is important to note that a species may occur in all the groups under analysis and still show a high IndVal, while its practical diagnostic value may remain questionable (see, e.g., species 8 in Table A3).

While the concentration notion seems to generalize the exclusiveness notion (i.e., a species considered exclusive to one vegetation unit has its presences fully concentrated in it), this is not entirely the case. We can easily think of a species with concentrated presences (or concentrated cover) in one vegetation unit, yet not exclusive to any of the vegetation units under consideration (see, e.g., species 5 and 8 in Table A3). All the previously mentioned indices (Bruelheide's *u*, IndVal, phi coefficient, the point-biserial correlation coefficient, and their variants) focus on the concentration without ensuring exclusiveness to any group(s). As shown, these indices can detect concentration of presences (or cover), but are not suitable for detecting differential species as originally defined and sought in phytosociological tabulation.

From Table A3, we can also conclude that $\mathrm{IndVal}^{g}_{ind}$, r^{g}_{ϕ}, r^{g}_{pb}, and DiffVal behave quite similarly for species that are absent from at least one group (species 1 to 4, 6 and 7), particularly when focusing on values for single groups. This suggests that these indices may behave similarly in broader vegetation analyses, spanning vegetation types that are floristically very different from one another. Although characteristic species were (and still are) considered unfeasible at the association level, this is not the case for higher syntaxonomic ranks, where high-fidelity species are found (Westhoff and van der Maarel 1978). As evidenced by the behaviour of these indices in such cases, the detection of this associated high fidelity holds irrespective of the paradigm.

On the significance testing for the association between a species and a group of relevés

Bruelheide (2000), Chytrý et al. (2002), and De Cáceres and Legendre (2009) focus on "statistical inference for species-site group associations". Nevertheless, while discussing this inference, De Cáceres and Legendre (2009, p. 3572) write:

> "*Finally, practitioners should keep in mind that if the classification of the sites has been obtained from the species composition itself (for example by K-means partitioning), the site groups would not be completely independent of the species data. In such a case of circularity, we can expect more significantly associated species than expected by chance only.*"

Table A3. Exemplifying how IndVal, the phi coefficient, the point-biserial correlation coefficient, and DiffVal behave depending on exclusiveness to groups, concentration of presences, and concentration of cover.

Relevé no. 0000000001|1111111112222|2222223 1234567890|1234567890123|4567890 **Group/ community** 1111111111|2222222222222|3333333	**Short description of the example:**	$\mathrm{IndVal}^{g}_{ind}$ for single groups 1|2|3	$\mathrm{IndVal}^{g}_{ind}$ for the best group combination	r^{g}_{ϕ} for single groups 1|2|3	r^{g}_{ϕ} for the best group combination	r^{g}_{pb} for single groups 1|2|3	r^{g}_{pb} for the best group combination	**DiffVal**
species 1 1111111111|0000000000000|0000000	Presences are concentrated in group 1 and exclusive to that group	1.00*|0.00|0.00	1.00* (group 1)	1.00*|-0.50|-0.50	1.00* (group 1)	1.00*|-0.50|-0.50	1.00* (group 1)	1.00
species 2 0000000000|0000000000000|1111111	Presences are concentrated in group 3 and exclusive to that group	0.00|0.00|1.00*	1.00* (group 3)	-0.50|-0.50|1.00*	1.00* (group 3)	-0.50|-0.50|1.00*	1.00* (group 3)	1.00
species 3 1110111010|0000000000000|0000000	Presences are concentrated in group 1 and exclusive to that group	0.70*|0.00|0.00	0.70* (group 1)	0.78*|-0.39|-0.39	0.78* (group 1)	0.78*|-0.39|-0.39	0.78* (group 1)	0.70
species 4 1110111010|0000000000000|1011110	Presences are concentrated in groups 1 and 3 and exclusive to those groups	0.35|0.00|0.36	0.71* (groups 1 and 3)	0.32|-0.67|0.34	0.67* (groups 1 and 3)	0.32|-0.67|0.34	0.67* (groups 1 and 3)	0.43
species 5 1110111010|0100010000100|0000100	Presences are concentrated in group 1 yet not exclusive to any group	0.46*|0.05|0.02	0.46 (group 1)	0.50*|-0.19|-0.32	0.50* (group 1)	0.50*|-0.19|-0.32	0.50* (group 1)	0.00
species 6 243035+534|0000000000000|0000000	Presences (and cover) are concentrated in group 1 and exclusive to that group	0.90*|0.00|0.00	0.90* (group 1)	0.93*|-0.46|-0.46	0.93* (group 1)	0.76*|-0.38|-0.38	0.76* (group 1)	0.90
species 7 243035+534|0000000000000|120211+	Presences (and cover) are concentrated in group 1 and 3 and exclusive to those groups	0.79*|0.00|0.10	0.88* (groups 1 and 3)	0.45*|-0.84|0.39	0.84* (groups 1 and 3)	0.72*|-0.44|-0.28	0.72* (group 1)	0.53
species 8 243035+534|000201000+00+|00020+0	Presences (and cover) are concentrated in group 1 yet not exclusive to any group	0.83*|0.01|0.01	0.83* (group 1)	0.57*|-0.27|-0.30	0.57* (group 1)	0.73*|-0.38|-0.36	0.73* (group 1)	0.00
species 9 3435355534|2232312121232|120211+	Cover is concentrated in group 1 (and group 2) yet not exclusive to any group	0.73*|0.20|0.06	0.97[1] (groups 1, 2, and 3)	0.16|0.16|-0.32	0.32 (groups 1 and 2)	0.84*|-0.27|-0.56	0.84* (group 1)	0.00
species 10 3435355534|2232312121232|12+211+	Cover is concentrated in group 1 (and group 2) yet not exclusive to any group. Presences in all relevés.	0.73*|0.20|0.07	1.00[1] (groups 1, 2, and 3)	NA|NA|NA	NA	0.84*|-0.28|-0.56	0.84* (group 1)	0.00

$\mathrm{IndVal}^{g}_{ind}$, r^{g}_{ϕ}, and r^{g}_{pb} were calculated using the "multipatt" function from the *indicspecies* package (De Cáceres and Legendre 2009). DiffVal was calculated using "tdv" function from the *diffval* package (Monteiro-Henriques and Cerdeira 2023). For the calculation of $\mathrm{IndVal}^{g}_{ind}$ and r^{g}_{pb}, the cover-abundance values of species 6 to 10 were transformed to percentage cover midpoints: 5 → 87.5%, 4 → 62.5%, 3 → 37.5%, 2 → 15%, 1 → 3%, and + → 1%. For species 1 to 5, $\mathrm{IndVal}^{g}_{ind}$ and r^{g}_{pb} correspond to their binary (presence/absence) versions (in fact, $r^{g}_{pb} = r^{g}_{\phi}$ for these species). $\mathrm{IndVal}^{g}_{ind}$ and r^{g}_{pb} use species 6 to 10 cover values in their calculation, whereas for the calculation of r^{g}_{ϕ} and DiffVal, cover-abundance values of species 6 to 10 were transformed into presence-absence values. |* The associated p-value was lower than 0.05, after 9,999 permutations (note that only the highest values of $\mathrm{IndVal}^{g}_{ind,}$ r^{g}_{ϕ} and r^{g}_{pb} are tested). |[1] p -value not computable in this case.

We must bear in mind that, in vegetation science, the classification of relevés is expected to emerge from species composition itself (i.e., be floristically based). Therefore, using such inference in this context is not advisable, as it would violate the independence assumption of significance testing (see Milligan 1996, pp. 366–367, and discussion of Example 1). In fact, De Cáceres et al. (2010) completely abandoned the concepts of fidelity and character species in their work and state on p. 1683:

> "*More indicator species will be found than expected by chance when the classification of sites has been obtained from the species composition itself (De Cáceres and Legendre 2009). In this case, p-values must be taken with caution: they do not result from a genuine test of significance since the classification of sites is not independent from the species data used in the indicator species analysis.*"

This is also why p-values are interpreted as a direct measure of fidelity rather than as probabilities in hypothesis testing (see, e.g., Tichý et al. 2010), often with lower thresholds than conventional significance levels. However, this approach significantly affects the method's ability to detect indicators (see Suppl. material 4).

For those interested in p-values, a time-consuming approach can be used to assess whether a given TDV (and the associated differential species pattern) could arise by chance – i.e., under the null hypothesis that the pattern does not reflect an underlying structure in species distribution. This can be tested with a Monte Carlo permutation test, optimizing TDV multiple times on randomized tables by permuting each species' presences among the relevés. A p-value is then obtained by computing the permutational probability, following Dufrêne and Legendre (1997), ranking the observed TDV within the distribution of randomly generated TDVs ordered in decreasing order.

Concluding remarks: Different contexts might call for different paradigms

The subtle changes found in Whittaker (1962) and Barkman (1989) bent the original concept of differential species. While these modifications may not be problematic in themselves, the differences – though seemingly minor refinements in the definition – lead to vastly different outcomes. This significantly restricts their application and usefulness for classification. Completely removing the subtlety of exclusiveness has consequences for the handling of vegetation data and has the potential to degrade both the implementation and the understanding of the tabulation technique.

The recognized loss of the exclusiveness notion of characteristic species at the association level should not have been extended to the concept of differential species. Floristics play a key role in the phytosociological approach. As it is currently postulated, concentration-based measures are ineffective in the vegetation classification process at the association level.

The DiffVal includes a measure of how exclusive a species is to one (or more) groups, i.e., a measure of the degree of exclusiveness. It is intended as a single measure for each species in a table, determined solely by the partition of the samples (note that, once a specific partition is given, it is straightforward to identify which species occur exclusively in each group or groups). In the case of IndVal, the phi coefficient, or the point-biserial correlation coefficient, comparisons are always pairwise (target group vs. all other groups). For these indices, it is irrelevant how the species' presences/absences are distributed among the remaining groups – i.e., outside the target group or combination of groups. This is not true for the DiffVal summands, marking a fundamental difference between DiffVal and the other indices.

I hope I have demonstrated why approaches based on the concentration notion are paradigmatically different from those based on the exclusiveness notion. This is the primary reason why comparisons between them are difficult, or even inconsistent. I believe that the block-based approaches are the only ones truly comparable to TDV-optimization.

Braun-Blanquet (1932, p. 62) presents an interesting reflection on the causes of fidelity. In brief, the author identifies several well-known natural phenomena, which are still considered today as underpinning general biogeographical and ecological patterns. Adapting to the terminology used today, the causes of fidelity are: (i) (paleo)biogeographical and environmental filtering; (ii) extreme-habitat adaptations; (iii) evolution/genetics; (iv) habitat availability; (v) biotic interactions (dependence); (vi) biotic interactions (competition). Centuries of research in botany, biogeography and ecology have shown that groups of species can be recognized as being associated with certain regions or environments throughout Earth's history. The seminal idea that vegetation communities could be defined floristically now has an inescapable (paleo)biogeographical and (paleo) ecological foundation.

In some regional contexts, particularly those with generally low diversity and low environmental roughness, an approach based on concentration without ensuring exclusiveness might be the only useful way of identifying vegetation types, as there may not be sufficient richness and turnover to find species exhibiting some degree of exclusiveness. In regions with high diversity, rich in endemism, and with significant environmental variation and roughness, an approach that ensures exclusiveness can be more effective for understanding the vegetation phenomenon, helping to reduce noise caused by numerous species with very high frequencies (e.g., dominant and companion species).

It is not my intention to raise any objections to approaches based on concentration. First, when analysing (or tabulating) highly dissimilar vegetation (e.g., higher syntaxonomic ranks), we have observed that the

concentration-based measures give similar results to those based on exclusiveness. Second, in regions with low species turnover and low environmental roughness, such approaches have proven useful.

The idea of a universal method to classify vegetation stands is appealing, but not necessarily feasible. Nearly a century ago, Braun-Blanquet (1932, p. 26) recognized this:

"The apparent simplicity of the analysis of the concept of vegetation is directly in contrast with the difficulty of making any universal rules. Sometimes it is entirely impossible to submit different vegetation types to similar methods of treatment."

The two different paradigms might have to coexist.

E-mail and ORCID

Tiago Monteiro-Henriques (Corresponding author, tmh@isa.ulisboa.pt), ORCID: https://orcid.org/0000-0002-4206-0699

Supplementary material

Supplementary material 1
R code for auxiliary functions (OptimClass and partition discrepancy) (*.pdf)
Link: https://doi.org/10.3897/VCS.140466.suppl1

Supplementary material 2
Annotated R code for constructing the Example 1 data set and partitioning via TDV-optimization (*.pdf)
Link: https://doi.org/10.3897/VCS.140466.suppl2

Supplementary material 3
Graphical output of the "tabulation" function for each partition obtained in Example 1, alongside ordination biplots (*.pdf)
Link: https://doi.org/10.3897/VCS.140466.suppl3

Supplementary material 4
Statistical power analysis of Fisher's exact test, adjusted for Example 1 (includes R code) (*.pdf)
Link: https://doi.org/10.3897/VCS.140466.suppl4

Supplementary material 5
Variation in OptimClass 1 with α for each clustering/partitioning method in Example 1 (*.pdf)
Link: https://doi.org/10.3897/VCS.140466.suppl5

Supplementary material 6
Analysis of local optima in the optimization procedures used in Example 1 (*.pdf)
Link: https://doi.org/10.3897/VCS.140466.suppl6

Supplementary material 7
Graphical output of the "tabulation" function for each partition obtained in Example 2, alongside ordination biplots (*.pdf)
Link: https://doi.org/10.3897/VCS.140466.suppl7

Supplementary material 8
Graphical output of the "tabulation" function for the first ten solutions of TDV-optimization, searching for three groups, in Example 2 (*.pdf)
Link: https://doi.org/10.3897/VCS.140466.suppl8

Supplementary material 9
Graphical output of the "tabulation" function for the first ten solutions of TDV-optimization, searching for two groups, in Example 2 (*.pdf)
Link: https://doi.org/10.3897/VCS.140466.suppl9

Supplementary material 10
A tentative bibliographic list of works containing numerical approaches for comparing or classifying vegetation plot data (*.pdf)
Link: https://doi.org/10.3897/VCS.140466.suppl10